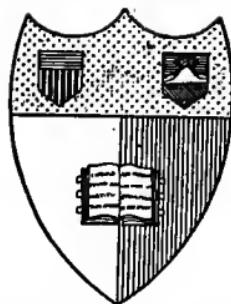


LB
3051
M75
E2



Cornell University Library
Ithaca, New York

FROM

Mrs. M.K. Cannon

Cornell University Library
LB3051.M75 E2

Educational tests and measurements,



3 1924 032 713 061
olin



Cornell University Library

The original of this book is in
the Cornell University Library.

There are no known copyright restrictions in
the United States on the use of the text.

RIVERSIDE TEXTBOOKS
IN EDUCATION

EDITED BY ELLWOOD P. CUBBERLEY
PROFESSOR OF EDUCATION
LELAND STANFORD JUNIOR UNIVERSITY

DIVISION OF SECONDARY EDUCATION
UNDER THE EDITORIAL DIRECTION
OF ALEXANDER INGLIS

ASSISTANT PROFESSOR OF EDUCATION
HARVARD UNIVERSITY

Walter W. Cannon

EDUCATIONAL TESTS AND MEASUREMENTS

BY

WALTER SCOTT MONROE, PH.D.

PROFESSOR OF SCHOOL ADMINISTRATION, AND DIRECTOR OF
THE BUREAU OF EDUCATIONAL MEASUREMENTS AND
STANDARDS, KANSAS STATE NORMAL SCHOOL

ASSISTED BY

JAMES CLARENCE DEVOSS, A.M.

ASSOCIATE PROFESSOR OF PSYCHOLOGY AND PHILOSOPHY
KANSAS STATE NORMAL SCHOOL

AND

FREDERICK JAMES KELLY, PH.D.

DEAN OF THE SCHOOL OF EDUCATION
UNIVERSITY OF KANSAS



HOUGHTON MIFFLIN COMPANY
BOSTON NEW YORK CHICAGO
The Riverside Press Cambridge

2-35-196
COPYRIGHT, 1917, BY W. S. MONROE, J. C. DEVOSS AND F. J. KELLY

ALL RIGHTS RESERVED

The Riverside Press
CAMBRIDGE · MASSACHUSETTS
U. S. A.

EDITOR'S INTRODUCTION

Up to very recently our chief method for determining the efficiency of a school system was the method of personal opinion. When the work of a superintendent of schools was called in question, the schools were visited and personal opinions expressed as to their standing. In case of a disagreement among the visitors the efficiency became a matter of dispute, and the people of a community usually favored the opinion which most nearly coincided with their prejudices and preconceived ideas.

Relatively recently the method of comparison was introduced. By means of this method the school system under consideration is compared with other school systems of the same size and class, and with reference to a number of different items. After such a comparison has been made, it is possible to place the school system relatively. If the school system studied stands fourth out of twenty school systems compared in one item, thirteenth in another, and at the bottom of the list in three others, it is not difficult to determine its position. It is evident that this is a much better method than the one of personal opinion. Its chief defect, though, lies in that the school system studied is continually compared with the average or median of its size and class. In other words, the school system is continually measured as against mediocrity, when as a matter of fact the average or median school system may not represent a good school

system at all. Perhaps all of the school systems below the average or median should be classed as poor school systems, and even some of those above are not doing what a school system should do.

Still more recently, and wholly within the past decade, a still better method for the evaluation of the work which teachers and schools are doing has been evolved. This new method consists in the setting up, through the medium of a series of carefully devised "Standardized Tests," of standard measurements and units of accomplishments for the determination of the kind and the amount of work which a school or a school system is doing. This new movement is as yet almost in its infancy, but so important is it in terms of the future of school administration that it already bids fair to change, in the course of time, the whole character of this professional service.

The significance of these new standards of measurement for our educational service is indeed large. Their use means nothing less than the ultimate transformation of school work from guesswork to scientific accuracy; the elimination of favoritism and politics from the work; the ending forever of the day when a personal or a political enemy of a superintendent can secure his removal, without regard to the efficiency of the school system he has built up; the substitution of well-trained experts as superintendents of schools for the old successful practitioners; and the changing of school supervision from a temporary or a political job, for which little or no preparation need be made, to that of a highly skilled piece of social engineering.

This new method for the evaluation of the work which a

school system is doing is so important that any young man or woman of to-day who desires to prepare for school administration should by all means thoroughly familiarize himself or herself with the aims and methods of this new type of administrative procedure. The underlying purpose of the new movement has been the creation of such standardized scales for measuring school work, and for comparing the accomplishments of different schools and groups of school-children, as to give to both supervisors and teachers definite aims in the imparting of instruction. Instead of continuing to teach without definite measuring-sticks, and to assign tasks and trust to luck and the growth process in children for results, which is comparable to the old-time luck-and-chance farming, it has been attempted to evolve standards of measurement which will do for education what has been done for agriculture as a result of the application of scientific knowledge and scientific methods to farming.

Such an important new movement is of especial significance to the teacher in charge of a class, to the citizen interested in schools, and to the superintendent responsible for results.

To the teacher it cannot help but eventually mean not only concise and definite statements as to what she is expected to do in the different subjects of the course of study, but the reduction of instruction to those items which can be proved to be of importance in preparation for intelligent living and future usefulness in life. It will mean, too, an ultimate differentiation in training for the different types of children with which teachers now have to deal, and the specialization of work so as to enable teachers to

obtain more satisfactory individual results. To the citizen the movement means the erection of standards of accomplishment which are definite, and by means of which he can judge for himself as to the efficiency of the schools he helps to support. For the superintendent it means the changing of school supervision from guesswork to scientific accuracy, and the establishment of standards of work by which he may defend what he is doing.

Up to the present time nearly all of the work which has been done in the evolution and testing out of these new standardized tests has been work of a highly scientific and technical nature, most of the articles being written in a language which the layman can scarcely understand. Often no interpretation has been attempted of the results which have been obtained. The classroom teacher and the school principal have naturally not found these studies of much help to them in their work.

This work has been carried far enough, however, so that the time now seems ripe for a clear and simple statement as to the nature of the different tests which have been evolved, their use, their reliability, what are the best standard scores so far arrived at, and, in particular, how to diagnose the results and apply remedial instruction. This the three authors of the present volume in the series have attempted to give, and, to make their work of the largest possible usefulness to normal-school students, teachers, and principals of schools, they have cast the whole in language so simple and untechnical that the average grade teacher can read the book and understand it. In addition, to give still larger value to the book, they have added a

number of chapters, written in a similar simple and readable style, giving the essential elements needed in understanding simple statistical methods, the meaning of scores, the unreliability of school marks and their relation to standardized scores, and the use of the standardized tests in the work of school supervision.

No space has been taken up in merely reproducing the tests themselves, though samples, showing their nature, have been inserted. If it is desired to use the tests with a class, they will be needed in quantities, and they may then be obtained in quantities and for very small sums from the persons and at the places mentioned in the chapter bibliographies. These bibliographies also give the most important book or article describing in detail the construction and use of the tests, in case the worker desires to go further than this volume presents. Instead, the authors have used their space in explaining to teachers and school officers the nature of the tests, telling how to give and score them, what standings the pupils should attain in their use, and presenting a rather full description as to the significance of the results obtained and how to remedy the defective conditions which the use of the tests reveals. In consequence, the book should prove of much use not only to students in normal schools and colleges, but to teachers and principals in our public schools as well. The style and contents of the volume are such as also to adapt it to reading-circle study with teachers, or to the needs of the average citizen interested in knowing something as to the nature and uses of the Standardized Tests.

ELLWOOD P. CUBBERLEY.

PREFACE

THIS book is designed primarily for teachers. It is based on two years' experience in giving a course on educational measurements to prospective teachers in a state normal school and on the experience received from directing a Bureau of Educational Measurements and Standards.

It is just twenty years since Rice startled the educators of this country by his proposal that the results of teaching spelling could be measured by a spelling test. His proposal was greeted with sarcasm and ridicule, but during the past two decades the opposition to the principle of educational measurements has almost entirely disappeared. To-day the widespread use of standardized tests and scales bears witness to the importance of this movement in American education. However, it is profitable to analyze our present interest in educational measurements. A thing may be interesting merely because it is new and spectacular. Scores are objective and are subject to graphical representation. A chart displayed attracts attention. Evidence is not wanting to show that a considerable number of teachers look upon educational measurements merely as an interesting topic for teachers' meetings or as a means of attracting attention in their community.

Standardized tests and scales are not "playthings." Neither are they teaching devices. They are instruments which furnish the teacher (1) with detailed and definite

aims, and (2) with a means for diagnosing the teaching situation which she faces. Unless the diagnosis is followed by remedial instruction the use of standardized tests and scales cannot be of much value. They become mere "play-things."

Our present tests are probably crude instruments, but the first railway locomotive was also crude. Even now standardized tests and scales are superior to ordinary examinations, but, more important, their use tends to engender in the teacher a type of thinking about her work which is very helpful. By using them she recognizes objective standards to be attained and not to be exceeded, the present achievements of her pupils, and that instruction must be suited to the needs of her pupils. When a teacher comes to think of her teaching problem in these terms, she is in a position to increase greatly her efficiency.

This book is addressed to teachers because they are charged with the instruction of pupils. The superintendent, principal, or student of education who is interested in the teacher's work also will find much of value in the book. Technical details of the derivation of tests are not given, but references are given so that one interested may pursue the matter. These were omitted because they are not essential to the use of the tests by teachers. For much the same reason the criticism of tests is made a secondary matter. The detailed criticism of tests and the derivation of improved ones must be left to the expert. The teacher needs to know only enough to enable her to choose wisely in selecting a test, and to prevent her from ascribing to the scores a significance which is not justified.

The newness of the field and the rapidity with which it is developing places limitations upon an attempt to write a text. It is recognized that probably before this volume is printed new tests will have been announced. However, the author believes that the point of view upon which the book is based is not merely temporary, and that, as new tests are available, the fundamental principles of the book may be applied to them.

It is obvious that in an endeavor such as this one must utilize the results obtained by many investigators. In fact it is hoped that this book may have the virtue of summarizing these results. The author is keenly aware of his obligation to all whose work is mentioned in the following pages. Special mention should be made of Professor DeVoss, who contributed the chapter on "Handwriting," and of Dean Kelly, who wrote the chapter on "Reading."

WALTER S. MONROE.

EMPORIA, KANSAS, April 27, 1917

CONTENTS

CHAPTER I. THE INACCURACY OF PRESENT SCHOOL MARKS

1

School marks — The inaccuracy of teachers' marks — Carter's investigation — Kelly's investigation — Conclusions from these studies — Johnson's investigation — Marking examination papers — Distribution of marks — Error due to unequal value of questions — Another example of unequal values — Rate of doing work neglected — Wide range of topics included within an examination — Most valuable topics for education — Questions and topics for investigation.

CHAPTER II. ARITHMETIC

17

I. THE PROBLEM OF MEASURING ARITHMETICAL ABILITIES.

Arithmetical abilities automatic or habits — Arithmetical abilities distinct — Separate types in handling integers — Each a specific habit — Why we need to use arithmetical tests.

II. STANDARDIZED TESTS FOR MEASURING ARITHMETICAL ABILITIES.

1. The Courtis Standard Research Tests, Series B — Marking the papers.
2. The Cleveland-Survey Arithmetic Tests — Spiral nature of these tests — Nature of the tests.
3. The Woody Arithmetic Scales — Measurement by means of a scale — The addition scale.
4. Research Tests in Addition of Fractions.
5. The Stone Reasoning Test.
6. Other reasoning tests.

III. STANDARD SCORES.

1. Courtis Standard Research Tests, Series B — Courtis Standard scores.
2. The Cleveland-Survey Tests — Cleveland and Grand Rapids scores.
3. The Woody Arithmetic Scales.
4. The Addition of Fractions Tests.
5. The Stone Reasoning Test.
The accuracy of individual scores — Gain or loss in repeating tests.

CONTENTS

IV. HOW TO HANDLE WHAT THE TESTS REVEAL.

Scientific management — Diagnosis of the teaching situation — Pupils' and class records charted — Meeting the situation; laws — Individual *vs.* class needs — Repeating the tests after an interval — Modifying the class drill — Use of practice tests — Questions and topics for investigation — Bibliography.

CHAPTER III. READING 66

THE PROBLEM OF MEASUREMENT IN READING.

I. SILENT READING.

(a) Recognition of words.

1. The Thorndike Visual Vocabulary Scales — Use and standards.
2. The Haggerty Visual Vocabulary Tests.
3. Starch's English Vocabulary Tests.

(b) Tests for comprehension and speed.

1. The Thorndike Scale Alpha.
2. The Minnesota Scale Beta.
3. The Courtis English Tests.
4. Brown's Silent Reading Test — Value of the test in diagnosis.
5. Starch's Silent Reading Tests.
6. Gray's Silent Reading Tests.
7. The Kansas Silent Reading Tests — Score value for the Kansas tests — Standard median scores.
8. The Courtis Silent Reading Tests.

II. ORAL READING.

1. The Jones Visual Vocabulary Tests.
2. The Haggerty Visual Vocabulary Tests.
3. Gray's Oral Reading Test — How the tests are scored.

III. AN ESTIMATE OF READING TESTS.

Criteria for estimating values — These criteria applied.

IV. THE SERVICE OF READING TESTS.

Service to the superintendent — Reveals wrong emphasis in teaching — Service to the teacher — Service to the child — Remedying the situation revealed — Types of situations revealed — A normal situation — To raise the median score — Overemphasis on oral reading — Care from the beginning — Reading above the primary grades — Reading in the upper grades — Where variability is too wide — The difficult but normal case; suggestions for helping — Questions and topics for investigation — Bibliography.

CONTENTS

CHAPTER IV. SPELLING	112
I. THE PROBLEM OF MEASUREMENT IN SPELLING.	
Difficulties encountered — The foundation words of the English language — Making a spelling test.	
II. SPELLING SCALES.	
1. The Ayres Spelling Scale — How constructed — Pupils who are not tested — How many words to use — Methods of giving the test — Letters per minute — What the Ayres scale really is — Directions for giving a timed sentence test.	
2. The Buckingham Spelling Scale.	
3. Starch's Spelling Scale — Measuring the extent of the ability to spell — Starch's spelling lists.	
III. STANDARDS.	
Ayres's scale — The Starch tests.	
IV. HOW TO LOCATE SPELLING DIFFICULTIES.	
Locating bad spellers — Individuality in spelling difficulties — "Spelling Demons" — Types of misspellings — Teaching the pupil to correct his errors in spelling — Causes of some misspellings — Good teaching of spelling — Devices for improving spelling — Making associations automatic — Courtis's spelling practice tests — Questions and topics for investigation — Bibliography.	
CHAPTER V. HANDWRITING	145
I. THE PROBLEM OF MEASUREMENT IN HANDWRITING.	
Another method of measuring.	
II. HANDWRITING SCALES.	
Measuring speed — Selections for the speed tests — Measuring quality ; use of scales — The score card for detailed analysis — The scales classified as to use — Methods of using scales — Measurement for diagnosis — Use of the score card — The Freeman Scale — Using the Freeman Scale.	
III. THE RELIABILITY OF MEASURES AND SCALES.	
Rate and quality contrasted — Accuracy of the scores — Training in using the scales — Relative values of the different scales.	
IV. STANDARD SCORES.	
Freeman's proposed standards — What these standards represent — Other evidence as to standards — Standards required for work.	
V. THE TEACHING SITUATION REVEALED.	
Handwriting a complex ability — An individual rather than a class problem — Children differ widely in abilities	

CONTENTS

and needs — Plotting scores, and reading their meaning — Successive measurements to reveal progress — Meeting the situation revealed — Systems of penmanship — Movement in handwriting — Rhythm — Speed — Quality and speed — General laws of learning applied — Devices of remedial instruction — Increasing speed — Developing rhythm — Motivating practice — Reasons for using handwriting scales — Questions and topics for investigation — Bibliography.

CHAPTER VI. LANGUAGE 192

I. THE PROBLEM OF MEASUREMENT IN LANGUAGE.

One of measuring specific habits.

II. THE MEASUREMENT OF ABILITY IN ENGLISH COMPOSITION.

1. The Hillegas Scale.
2. The Harvard-Newton Composition Scale.
3. The Breed and Frostic Scale.
4. Willing's Scale.
5. The Nassau County Supplement.

Reliability of measurements — Use of the scales — Directions for using the Hillegas Scale — Hillegas Scale scores — Directions for using the Harvard-Newton Scale — The Harvard-Newton Scale scores — Directions for using the Willing Scale — The Willing Scale scores — The Willing Scale reproduced.

III. THE MEASUREMENT OF LANGUAGE ABILITY BY COMPLETION TESTS.

The Trabue Completion-Test Language Scales — Trabue test standards.

IV. THE MEASUREMENT OF ABILITY IN ENGLISH GRAMMAR.

Types of ability — Starch's Grammatical Scales — The Punctuation Scale — The Grammar Tests — Buckingham's Test.

V. MEASURING ACCURACY IN COPYING.

The Boston test — Kinds of errors made — Misspelled words — Undotted "i's" and uncrossed "t's."

VI. EDUCATIONAL SIGNIFICANCE OF THE USE OF THESE SCALES AND TESTS.

Finding specific language weaknesses — Remedyng the situation revealed — Analyzing language ability — Questions and topics for investigation — Bibliography.

CHAPTER VII. HIGH-SCHOOL SUBJECTS 224

I. ALGEBRA.

The problem of measurement — The fundamental opera-

CONTENTS

tions of Algebra — Standard Research Tests in Algebra — Other Algebra tests — Conclusions from the tests — Standards — Meeting the teaching situation revealed by Algebra tests.	
II. GEOMETRY.	
III. FOREIGN LANGUAGES.	
Starch's language tests — The Hanus Latin Tests. Henmon's Latin Tests.	
IV. PHYSICS.	
V. OTHER TESTS WHICH MAY BE USED IN THE HIGH SCHOOL.	
Questions and topics for investigation — Bibliography.	
CHAPTER VIII. STATISTICAL METHODS	241
Good arrangement of scores — The median — Frequency of scores — Intervals of distribution — Approximate and true median — The average — The mode — Measures of variability: (1) Average deviation; (2) Percentiles; quartiles; probable error — Correlation — Coefficient of correlation — Graphical representation — Representing three qualities — Representing many qualities — Questions and topics for investigation.	
CHAPTER IX. THE MEANING OF SCORES	258
Indefiniteness of school marks — School marks <i>vs.</i> scores — Translating school marks into scores — A satisfactory standard — An efficient standard — Effort to be expended on the tool subjects — School demands on the tool subjects — Basis for standards of accomplishment — Types of standards — Questions and topics for investigation.	
CHAPTER X. THE DERIVATION OF TESTS, AND EX- AMINATIONS	273
Analyzing pupils' class work — Bases for evaluating pupils' work — The cycle principle — The per-cent-of-pupils-solving basis — A normal distribution of ability — Points to be considered in evaluating exercises — Opinion of competent judges — The teacher-judgment basis — Reliability important.	
Making examinations more effective — Care in framing questions — Questions and topics for investigation.	
CHAPTER XI. USE OF STANDARD TESTS IN THE SUPER- VISION OF INSTRUCTION	284
Assisting teachers — Four steps in supervising instruction with tests — Giving the tests — Tabulating the scores — Interpretation of the scores — Remedial treatment — Teachers need detailed and definite specifications — Courses of study represent working	

CONTENTS

specifications — Such subject-matter directions not quantitative — Such directions lead to formal and uniform instruction — The Tests aim to introduce quantitative work — Tests introduce scientific management — Handwriting an example of wasting time — The Cleveland Reading results a study in efficiency — Standards for instruction illustrated from Arithmetic — Results of using the Courtis tests in Boston — The supervisor and the standard tests — Questions and topics for investigation.

LIST OF CHARTS AND FIGURES IN THE TEXT

1. Distribution of marks assigned to one Geometry paper by 116 teachers	7
2. Distribution of scores with the Stone Reasoning Test in Butte	43
3. A chart, showing the scores made by a sixth-grade pupil, in comparison with standard scores, using the Courtis Arithmetic Tests	51
4. Median scores for all Cleveland schools, and five selected schools (Test D, division)	52
5. The Ayres Spelling Scale	113
6. Showing distribution of 91 pupils, according to the number of words spelled correctly	117
7. A Section of the Thorndike Handwriting Scale	149
8. Two sections of the Ayres Handwriting Scale	150-151
9. Standard score card for measuring handwriting	155
10. Individual record card, Freeman Scale	160
11. Showing the distribution of scores in handwriting of a third-grade class	177
12. Showing the distribution of scores in handwriting of a fourth-grade class	177
13. Showing the distribution of scores in handwriting of a fifth-grade class	177
14. Results of the Composition Test in Salt Lake City	201
15. Showing correlation of grades at entrance to college and in the Freshman year	253
16. Graphic representation of the standard scores for Starch's Spelling Tests	254
17. Another form of graphic representation of the same standard scores as in Figure 16	254
18. Representing graphically the standard for handwriting	255
19. Representing a pupil's score in several subjects	256
20. Showing the meaning of school grades in terms of scores	261

LIST OF CHARTS AND FIGURES

21. Showing a normal distribution of pupils according to ability	276
22. Distribution in rank of 47 cities, arranged in classes according to time spent on handwriting	295
23. Average scores in speed and quality of silent reading in each grade in Cleveland, and in 13 other cities	297
24. Same, in three selected Cleveland schools	298

EDUCATIONAL TESTS AND MEASUREMENTS

CHAPTER I

THE INACCURACY OF PRESENT SCHOOL MARKS

School marks. Educational measurements are not new in school work. Since schools have existed, teachers and other school officials have attempted to measure the abilities of pupils by estimating daily recitations and by examinations. The measures of the abilities of pupils obtained in these ways are thought to possess a high degree of precision and are treated very seriously.

The promotion of pupils depends upon the "grades" they receive. The ability of a pupil in each of the subjects is measured by the teacher's estimate and by examination, and, if the resulting measures show the pupil to be a few points, or in some instances a fraction of a point below the "passing mark," the pupil is classified as a failure. If the resulting measures equal or are above the "passing mark," the pupil is promoted.

The "grades" or school marks are entered upon the monthly or quarterly report cards. Parents, as well as teachers and pupils, take these school marks very seriously. If Johnnie's "grades" for a given month are below those of the preceding months, or, worse still, if they are below those of neighbor Smith's Mary, an explanation is demanded. A

2 EDUCATIONAL TESTS AND MEASUREMENTS

permanent record is kept of at least the yearly "grades," and the awarding of school honors is based upon it.

Until recently, practically all admission to college was determined by examination. Except in the universities and colleges of the Central and Western American States the custom still maintains generally throughout the world. This practice is based on the assumption that the examining committee can determine thereby the effectiveness of the candidate's college preparatory work. The civil service, from its inception in China centuries ago until the present day, has employed the examination as a means for measuring the ability of persons who desire positions operated under this system.

The inaccuracy of teachers' marks. Within the last few years a number of investigations have been made to ascertain the accuracy or reliability of measures obtained by means of teachers' estimates and by means of examinations. In the world of physical things we measure distance by means of the yardstick, mass by means of scales, the volume of liquids by means of gallon measures. Measures of these magnitudes, when made carefully with accurate instruments, possess a high degree of reliability. By a high degree of reliability we mean, for example, that if two persons measure the length of the same room by means of the same yardstick or any other yardstick, the two measurements will be approximately equal. If they differ by more than one or two inches, we doubt the accuracy of both, and we demand that the room be measured again. Similarly, in the case of school-children, if we find that when the same children are measured in the same subjects by two different teachers, the two sets of measures

do not agree rather closely, we have reason to doubt the accuracy of both sets of measures. On the other hand, if the two sets of measures ("grades") agree closely, we have reason to believe them accurate or reliable.

Carter's investigation. In 1911, Carter¹ investigated the school marks which had been given to the pupils who completed the eighth grade in three elementary schools in the city of Milwaukee, Wisconsin, and who entered the same high school. It was found in the case of arithmetic that two thirds of the marks given in school B were below 78; in school A, one third were below 79; in school C, one third were below 82. Taking the higher marks, in school B, one third were above 78; in school A, two thirds were above 79, and one third above 84; in school C, two thirds were above 82 and one third above 88. According to these marks it is evident that the pupils in school C received much higher marks than did the pupils in the other two schools, and that the pupils of school B were judged to be conspicuously inferior to the pupils in the other schools. If these marks represent accurate measures of the ability of these pupils in the field of arithmetic, we would expect the pupils in school C to receive the highest marks in mathematics when the pupils from the three schools went to the same high school.

Carter sums up his conclusions in these words: "When the rank of the pupils in arithmetic was compared with their rank in algebra, it was found that a greater percentage of school B (the school which gave the lowest marks) excelled in maintaining their original rank or increasing it. In fact,

¹ Carter, R. E., "Correlation of Elementary Schools and High Schools"; in *Elementary School Teacher*, vol. 12, pp. 109-18.

4 EDUCATIONAL TESTS AND MEASUREMENTS

there was a complete reversal of things from what the absolute marks might indicate." Thus, we find that the two sets of measurements of these pupils are characterized by differences rather than agreement. We, therefore, have the evidence of this investigation that the marks assigned by the teachers were inaccurate measures of the abilities of the pupils.

Kelly's investigation. In 1913, Kelly¹ made a similar investigation of the marks given to the sixth-grade pupils in four ward schools in Hackensack, New Jersey, and the marks given to the same pupils when they went to a common departmental school for seventh-grade work. He states his conclusions as follows: "This means that for work which the teacher in school C (one of the ward schools) would give a mark of 'G' (good) in language, penmanship, or history, the teacher in school D (another ward school) would give less than a mark of 'F' (fair)."

Conclusions from these studies. From these two investigations (and many others have been made which might be quoted²) it is clear that when different teachers measure the abilities of the same pupils in the same subjects by means of examinations and estimates of recitations, they give different "grades." Hence, we must conclude that teachers' marks are unreliable, that is, they are in general inaccurate measures of the abilities of pupils.

Johnson's investigation. Another type of investigation has been made by Johnson,³ Principal of the University

¹ Kelly, F. J. *Teachers' Marks*. (*Teachers College Contributions to Education*, no. 66, p. 7.)

² See Kelly, F. J. *Teachers' Marks*, pp. 6-50.

³ Johnson, F. W. "A Study of High School Grades"; in *School Review*,

High School of the University of Chicago. In the University High School, "F" denotes failure, and the four successive ranks above failure are indicated by "D," "C," "B," and "A." For the several departments of the school, Johnson tabulated the number of times each mark was given during the years 1907-08 and 1908-09. The facts revealed by these tabulations may be illustrated by the following. In English the per cent of failures was 15.5, which was nearly double that of history (8.1). The highest mark ("A") was given to 9.3 per cent of the pupils taking French, but in the German department it was awarded to 17.1 per cent. English and history occupied similar places in the program of studies. They were taken by practically all students. French and German likewise occupied similar places in the school. Thus, there is no apparent reason why these differences in marking pupils should exist. This lack of uniformity indicates the lack of uniform standards in marking pupils which, of course, means that the marking is probably inaccurate.

Marking examination papers. The written examination is the most common means of measuring the abilities of pupils, although many teachers and school patrons oppose its use. They contend that pupils working under pressure frequently become nervous and confused and consequently cannot do themselves justice, while other pupils, who have no real grasp of the subject, are able by cramming to write excellent papers. It is also contended that the questions are frequently not well selected and do not pertain to the essentials of the subject.

vol. 19, pp. 13-24. See also Kelly, F. J., *Teachers' Marks*, p. 11, and following for reports of similar investigations.

6 EDUCATIONAL TESTS AND MEASUREMENTS

There is probably some truth in the above assertions, but within the past few years there have been a number of investigations to ascertain if teachers mark examination papers accurately, assuming that what appears on the papers is a true record of the abilities of the pupils. Starch and Elliott¹ investigated the accuracy with which teachers marked papers in English, geometry, and history. Their method and the facts revealed may be illustrated by the case of geometry.

A facsimile reproduction was made of an actual examination paper in plane geometry. A copy of this reproduction was sent to each of the high schools included in the North Central Association of Colleges and Secondary Schools, with the request that it be marked on the scale of one hundred per cent by the teacher of geometry. Papers were returned from 116 schools, and the results tabulated. When we consider that the subject-matter of geometry is quite definite, and that the papers were marked by teachers who were thoroughly acquainted with the subject, it would seem that we might expect the marks or "grades" placed upon this examination paper to be in close agreement. However, exactly the opposite was the case.

Distribution of marks. The distribution of the marks is shown in Fig. 1. Of the 116 marks, two were above 90, while one was below 30. Twenty were 80 or above, while twenty other marks were below 60. Forty-seven teachers

¹ Starch and Elliott. "Reliability of Grading High-School Work in English"; in *School Review*, vol. 20, pp. 442-57: "Reliability of Grading Work in Mathematics"; in *School Review*, vol. 21, pp. 254-59: "Reliability of Grading Work in History"; in *School Review*, vol. 21, pp. 676-81.

assigned a mark passing or above, while 69 teachers thought the paper not worthy of a passing mark.¹

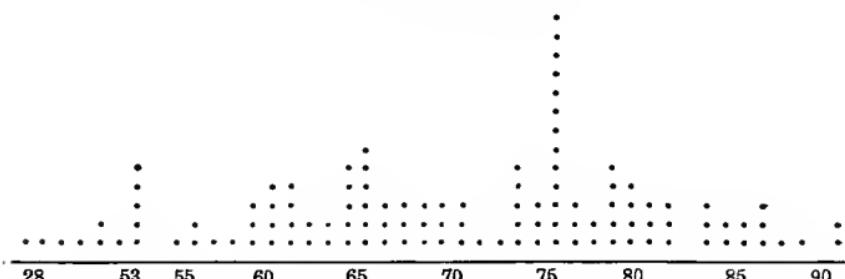


FIG. 1. DISTRIBUTION OF MARKS ASSIGNED TO ONE GEOMETRY PAPER BY 116 TEACHERS

Passing grade 75. Range 28 to 92. Marks assigned by schools whose passing grade was 70 were weighted by 3 points. Median 70. Probable error 7.5

Not only were similar results obtained by Starch and Elliott in English and in history, but other investigators² have verified them many times. In the face of such facts only one conclusion is possible; namely, that under ordinary conditions the marks assigned to examination papers by teachers are very unreliable. Such marks can represent only very crude and very inaccurate measures of the abilities of pupils. It is not too much to say that the mark which a pupil receives on an examination paper depends upon the teacher who grades the paper, as well as upon what the pupil places upon the paper.

It has also been shown that the same teacher is not consistent in his own marking. If a set of papers are marked a second time, the two sets of marks will vary widely.³

¹ See Starch, Daniel. *Educational Measurements*, p. 8, for the factors which are responsible for this inaccuracy.

² See Kelly, F. J. *Teachers' Marks*, p. 51, and following, for accounts of other investigations.

³ See Starch, Daniel. *Educational Measurements*, p. 9.

8 EDUCATIONAL TESTS AND MEASUREMENTS

Error due to unequal value of questions. A critical study of examinations and of the manner of giving them reveals other causes of inaccuracy in teachers' marks. In the first place, the questions are generally considered equal in value, and a pupil is given as much credit for answering a very easy question as for answering a difficult one. Or, if the questions are not considered equal in value, values are arbitrarily assigned by the teacher upon the basis of her estimate of the importance of the questions rather than upon the basis of their difficulty for pupils. In the case of the most fundamental facts, the importance may be used as the criterion of value, but in general the difficulty of the questions and the time required for answering them are used as the criteria of value. The difficulty of a question is represented by the per cent of correct responses.

Investigation has shown that teachers' estimates of the values of examination questions vary as widely as do the marks which they assign to examination papers. Inglis¹ sent a set of ten questions in plane geometry to about three hundred teachers of mathematics in the high schools of the Middle States and New England, with the request that each teacher assign to each question the number of points which should be allowed for a correct answer. The only limitation placed upon the teachers was that the total number of points should equal 100. Out of 122 replies which were received, 20 were to the effect that a value of ten points should be assigned to each question. The variations of the remaining 102 judgments are shown in Table I.

¹ Inglis, Alexander. "Variability of Judgments in Equalizing Values in Grading"; in *Educational Administration and Supervision*, vol. 2, pp. 25-30.

It is interesting to note that the values 5, 8, 10, 12, and 15 were used most frequently. This was probably because they are convenient values and because the teachers were accustomed to use them and not because of any scientific determination of the value of the questions. The range of values is surprisingly large. That teachers should differ so widely in their judgments is significant. It emphasizes the chaotic condition which now exists and will continue to exist until we have standardized tests.

TABLE I. THE VARIABILITY OF TEACHERS' ESTIMATES OF THE VALUE OF EXAMINATION QUESTIONS

Value assigned (per cent)	Number of judges assigned value to each question											Per cent
	1	2	3	4	5	6	7	8	9	10	Total	
0.....	1	1	0.1
1.....	1	1	0.1
2.....	..	1	1	1	1	4	0.4
3.....	1	..	1	2	0.2
4.....	..	3	1	5	0.5
5.....	14	15	7	1	16	2	8	6	8	19	96	8.8
6.....	4	7	1	1	6	..	2	1	..	13	35	3.4
7.....	6	3	1	3	6	4	6	2	4	8	43	4.2
8.....	23	26	15	14	29	6	14	8	7	18	160	15.7
8.5.....	..	1	1	1	1	..	4	0.4
9.....	3	5	5	4	4	..	3	3	2	2	31	3.0
10.....	34	30	44	24	27	34	31	24	16	18	282	27.6
11.....	2	3	4	6	1	3	2	2	4	1	28	2.8
11.5.....	1	1	1	1	..	4	0.4
12.....	8	6	16	23	6	21	16	24	21	10	151	14.8
12.5.....	..	1	1	2	2	2	..	8	0.8
13.....	3	..	2	2	2	3	5	2	2	1	22	2.2
14.....	1	..	1	3	1	1	1	..	8	0.8
15.....	3	1	5	16	4	18	9	21	26	7	110	10.9
16.....	2	..	2	..	1	3	1	9	0.9
17.....	1	1	0.1
18.....	1	1	1	1	2	1	7	0.7
19.....	0.0
20.....	3	..	2	1	1	7	0.7
25.....	1	1	0.1
Range.....	12	13	15	20	14	20	13	16	18	18	25	
Average.....	8.9	8.3	10.0	11.1	8.5	11.5	10.1	11.1	11.6	8.6	10.0	
A. D. (Av.).....	2.0	1.9	1.6	2.5	2.0	2.3	2.3	2.5	2.8	2.8	2.4	
Median.....	9.0	7.5	9.5	10.7	7.8	11.5	9.6	11.1	11.3	7.6		
Med. Dev. (M).....	.8	1.8	.6	1.2	1.3	1.2	1.6	1.1	2.8	2.0		

The terms "average," "average deviation," "median," and "median deviation," are defined in Chapter VIII.

10 EDUCATIONAL TESTS AND MEASUREMENTS

When the questions were given to a class of about forty high-school students, their answers to the questions indicate the following relative difficulties of the questions:—

Question number 1	2	3	4	5	6	7	8	9	10	
Total credits...	187	191	135	102	260	8	20	15	0	298

In scoring the papers credit was given for answers partly right. Since question 10 has the largest number of credits, it is shown to be the least difficult, and therefore to have the least value. Question 9 has the greatest value. The other questions rank between these two in value.¹

Another example of unequal values. The difference in the values of questions is illustrated by the results of giving the following examination in arithmetic to a sixth-grade class. Out of a class of 31 pupils, 17 answered the first question correctly, 29 the second, 12 the third, and 20 the fourth. It is very evident that for the pupils of this particular class the second question was the easiest of the four. If the second question is considered as having a value of 10 points, certainly the other three questions should have higher values.

1. Write in Roman system: 49, 79, 94, 96, 146.
2. If 11 A. of land are worth \$1485, what is one acre worth?
3. If a desk is $4\frac{3}{4}$ ft. long and $3\frac{5}{8}$ ft. wide, what is the perimeter?
4. How much must you add to $26\frac{7}{8}$ in. to make a yard?
5. A man has to travel 117 mi. After going $\frac{5}{9}$ of the distance, how many miles has he still to travel?
6. The perimeter of a square is 851 in. What is the length of one side?
7. Of 152 chickens a hawk captured $12\frac{1}{2}\%$. How many were captured? How many were left?
8. A man saves \$675.20 a yr., which is 32% of his income. How much is his income?

¹ See page 115 for similar results on the reliability of teachers' judgments of the relative difficulty of the spelling of words.

9. At \$1.38 a yd., what will 37 yds. of carpet cost?
10. At \$65.50 an acre, what must a man pay for 25.4 acres of land?

It is easy to understand how a serious element of error is introduced when each question is considered to have a value of 10 points and the questions are not equal in difficulty. The situation is much the same as we would have in measuring distances if yardsticks of different lengths were used, but were considered to be equal. Under such circumstances, a yard would have no definite length, and to say that a certain distance was 21.42 yards would convey no definite information about it. For this reason the Federal Government has standardized all weights and measures by establishing definite units, and before we can obtain definite measures of the abilities of children, it will be necessary to devise tests consisting of standard units.

Rate of doing work neglected. In the second place, it is customary in giving an examination to allow sufficient time for all pupils to answer all of the questions, or if this is not done, the papers are graded on the basis of what each pupil has done. This manner of giving an examination fails to take into account the rate at which a pupil is able to answer the questions. Only the quality of the answers is considered, and the pupil who answers the questions with difficulty and who barely finishes in the time allowed, receives exactly the same "grade" as the more capable pupil who is able to answer the questions easily and who finishes in one half or one third of the time, providing the two sets of answers are equivalent. It is clear that when this is done, the "grade" or mark which the pupil receives is not a true

12 EDUCATIONAL TESTS AND MEASUREMENTS

measure of his ability, because the rate at which he is able to do work is just as much a factor of his ability as is the quality of what he does.

Some may insist that it is unfair to the slow-working pupil not to allow sufficient time for him to answer all of the questions. However this may be, it certainly is unjust to the more capable pupil to deprive him of the opportunity to demonstrate what he is able to do. This is exactly the case when the work asked of him is sufficient to keep him employed only a half or a third of the period allowed for the examination. This practice of ignoring the rate of working probably tends to cause desultory and careless school work.

Investigation has shown that rapid work and a high degree of quality or accuracy are not incompatible in arithmetic. The same statement could probably be made with reference to reading. Investigation has indicated that a considerable per cent of pupils can be made more accurate in arithmetic by forcing them to work more rapidly. It has also been shown that about three pupils out of four make progress in speed and accuracy at the same time. In view of these facts, it appears that good instruction requires that the teacher give attention to the rate of doing work as well as to the quality of the work done. The rate at which a pupil is able to do work of a given quality is as much a factor of his ability as is the quality of the work which he does.

The rate at which a pupil works can be measured very easily. It is simply necessary to secure a record of the time which he spends in answering the set of questions. When an examination is given to a group, it is rather inconvenient to secure a record of the time which each pupil spends upon

the examination. However, one can secure just as true a record of the rate at which each pupil works by making the examination long enough so that no pupil finishes in the time allowed. For each pupil the number of minutes, divided by the number of units of work which he did, will give his rate of working per unit.

Wide range of topics included within an examination. In the third place, examinations are usually made up of questions from a number of different fields within a subject. Take, for example, the questions given on page 10. Question 1 calls for a knowledge of Roman numerals; question 2 asks the pupil to find the cost of a unit when the cost of the whole is given; questions 3, 4, and 6 deal with mensuration; question 5 calls for the finding of a fractional part of the whole; questions 7 and 8 are problems in buying. Thus we find six different topics included within an examination of ten questions.

Suppose a pupil receives a "grade" of 80 on this examination. Even if 80 is an accurate measure of what the pupil is able to do on this examination, it cannot have a definite meaning. It does not tell us whether the pupil lacks ability in the field of Roman numerals, or in the field of percentage, or in some other of the fields included in this examination. In order that the total score made on an examination may be a definite measure of a pupil's ability, the questions which compose it must be drawn from a single field, or at most from a small group of closely related fields. If this is not done, the scores for each question must be kept separate.

The situation is much the same as if the length, width, height, seating capacity, number of windows, and the num-

14 EDUCATIONAL TESTS AND MEASUREMENTS

ber of doors of a room were added together to form a measure of the room. If we assume that each of these characteristics of the room was measured with a high degree of accuracy, the total of the numbers expressing the measures gives us only very general information about the room. If the total is large, we know that the room is probably large; if the total is small, we know that it is small. But under no circumstances can we be certain that the room has any windows or doors, that it contains any seats, or that its dimensions are well proportioned. In order that we may have definite information about the room, it is necessary that the measures of the several characteristics be kept separate.

It is obvious that the questions of an examination should pertain to the significant topics of a subject if the examination is to furnish valuable information. In the case of measuring a room, a number of characteristics of the room might be measured. For example, one might measure the diagonals of the floor and walls, the height of the chairs, the color of the walls, the quality of the finish, etc. Such measures would be important for certain purposes, but if our purpose is to learn about the size of the room they are not very significant, and, hence, are not valuable. For this purpose the valuable measures are the width, length, and height of the room. If we have another purpose, for example, to determine the quality of the lighting of the room, other measures are the valuable ones.

Most valuable topics for education. As yet only a few studies have been made to determine the topics within our school subjects which are most valuable for the purpose of

education. Ayers¹ has determined the 1000 words which are used most frequently in writing and, hence, whose spelling is the most valuable.² Charters³ has determined the rules of grammar which the children of Kansas City, Missouri, need to learn in order to correct the errors in their language, both oral and written. Freeman⁴ has analyzed handwriting into its significant factors.⁵ A systematic attempt is being made by a Committee of the Department of Superintendence of the National Education Association to determine the minimum essentials of the common branches. The reports of this committee have appeared as the Fourteenth Yearbook,⁶ part 1, and the Sixteenth Yearbook,⁷ part 1, of the National Society for the Study of Education.

¹ Ayers, L. P. *Measurement of Ability in Spelling*. (Russell Sage Foundation Bulletin.)

² See chapter IV.

³ Charters, W. W. and Miller, Edith. *A Course of Study in Grammar based upon the Grammatical Errors of School Children of Kansas City, Missouri*. (University of Missouri Bulletin, vol. 16, no. 2.)

⁴ Freeman, F. N. *The Teaching of Handwriting*. (Houghton Mifflin Company, Boston, 1914.)

⁵ See chapter V.

⁶ *Minimum Essentials in Elementary-School Subjects—Standards and Current Practices*. (152 pp. 1915.)

⁷ *Second Report of the Committee on Minimal Essentials in Elementary-School Subjects*. (192 pp. 1917.)

16 EDUCATIONAL TESTS AND MEASUREMENTS

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What should be the purpose of examinations? Does the fact that examination papers are not marked accurately mean that examinations should not be given?
2. In view of the proven inaccuracy of teachers' marks, is a difference of less than five per cent between "grades" significant?
3. In life outside of school is the rate at which one is able to work considered?
4. Which gives the truer measure of the ability of all pupils; (a) an examination of definite length for which unlimited time is allowed, or (b) an examination for which the time is limited so that no one finishes?
5. Should the fact that a pupil fails to complete an examination in the time allowed be considered a reason for raising or lowering the mark given to that pupil's paper?
6. Have a set of examination papers marked independently by several teachers and compare the sets of marks. If possible secure from the teachers reasons for their differences of opinion.
7. Make a study of the distribution of marks given to the same pupils by high-school teachers.
8. Secure a set of examination papers written by pupils whose handwriting you do not know. Mark the papers recording your mark on a separate sheet. Do this several times at intervals of a week or ten days. Then compare the several sets of marks.
9. Why should "catch" questions and irrelevant questions be condemned?

CHAPTER II

ARITHMETIC

I. THE PROBLEM OF MEASURING ARITHMETICAL ABILITIES

IN measuring a physical object, such as a room, chair, haystack, or irregular-shaped field, it is necessary first to determine what dimensions are to be measured. For example, if our purpose is to ascertain the number of yards of carpet needed to cover a floor, only certain characteristics of the room, namely, length and width, are significant. On the other hand, if our purpose is to obtain a numerical measure of the lighting of the room, other characteristics, such as the number, position, and area of the windows, are essential.

It is obvious that, before our efforts to make educational measurements can be intelligently directed, we must know what we are to measure, and what the significant features are. Thus, the first step is to analyze the outcomes of instruction and to determine the significant characteristics of the elemental abilities.

Numerous arithmetical outcomes have been recognized in the statements of the aim of teaching arithmetic, but it is generally agreed that among the desired outcomes of arithmetical instruction are the abilities required to perform the operations of addition, subtraction, multiplication, and division with integers and with fractions, both common and decimal.

Arithmetical abilities automatic or habits. The pupil must be able to perform the operations of arithmetic rapidly and with a minimum of attention if he is efficient. As soon as he recognizes that a multiplication combination is called for, as, for example, 8×7 , the response, 56, must be forthcoming immediately. Time cannot be taken to think out the product. The pupil's attention must be reserved for deciding what operations to perform in dealing with the problems of arithmetic. The situation is similar to that which we have in any field of action where particular acts occur frequently and are always the same. Such acts must be reduced to the plane of habit if a person becomes skillful. We may therefore describe these arithmetical abilities as being automatic or habits.

Arithmetical abilities distinct. A few years ago Stone¹ investigated the nature of ability in arithmetic and concluded that it was made up of a number of specific abilities. His conclusions have been corroborated by a number of other investigations,² and it is now reasonably certain that in teaching the operations of arithmetic, we are attempting to engender a number of specific abilities which are relatively distinct, and not a single arithmetical ability. There are as many different abilities as there are types of examples. In fact, it is obvious that the ability to add a column of three

¹ Stone, C. W. *Arithmetical Abilities and Some Factors Determining Them.* (Teachers College Contributions to Education, no. 19. 1908.)

² Ballou, F. W. *Determining the Achievement of Pupils in Addition of Fractions.* (School Document no. 3, 1916. Boston Public Schools.)

Recently an investigation was made, under the direction of the writer, of the nature of the ability to place the decimal point in a quotient. This investigation showed that a number of specific abilities were involved, and not a single ability.

figures is not the same as the ability to add a column of twelve figures. In adding a column of figures it is necessary that one hold in mind the partial sum until he has added the next figure. This process must be repeated continuously until the final sum is reached, and a failure to do this continuously will result in stopping the adding, at least temporarily. It is a frequent occurrence, for one who is not accustomed to adding long columns of figures, to find that he has stopped, perhaps has even lost the partial sum, and must begin again. The span of attention required in adding three figures is short, and pupils who are able to do examples of this type with a high degree of skill frequently are unable to add long columns of figures with an equal degree of skill. In fact, we have no reason to expect them to be able to do this type of example until they have practiced upon it.

Separate types in handling integers. Courtis,¹ the author of the Standard Research Tests, has identified the following types of examples in the operations with integers:—

Addition: (1) addition combinations; (2) single-column addition of three figures each ; (3) “bridging the tens,” as $38 + 7$; (4) column addition, seven figures; (5) carrying; (6) column addition with increased attention span, thirteen figures to the column; (7) addition of numbers of different lengths.

Subtraction: (1) subtraction combinations; (2) subtraction of 9 or less from a number of two digits, both with and without simple “borrowing”; (3) subtraction involving borrowing.

¹ Courtis, S. A. *Teacher's Manual for Courtis Standard Practice Tests* (1916).

Multiplication: (1) multiplication combinations; (2) multiplicand two digits, multiplier one digit, and no carrying; (3) same as number 2, but with carrying; (4) long multiplication, without carrying; (5) zero difficulties; (6) long multiplication, with carrying.

Division: (1) division combinations; (2) simple division, no carrying; (3) same as number 2, but with carrying; (4) long division, no carrying; (5) zero difficulties, without carrying; (6) long division, with carrying, "first case," the first figure of the divisor is the trial divisor and the trial quotient is the true quotient; (7) "second case, where the trial divisor is one larger than the first figure of the divisor, but the trial quotient is the true quotient"; (8) "third case, where the first figure of the divisor is the trial divisor, but the true quotient is one smaller than the trial quotient"; (9) "fourth case, where the first figure of the divisor must be increased by one to get the true quotient."

Each a specific habit. Each of these types of examples requires a specific habit or automatism. To be sure, certain elements, such as the fundamental combinations, are common elements, but careful analysis will show that the ability to do examples of one type is different from that required to do another. Not only will a careful analysis reveal this fact, but it has been repeatedly demonstrated by carefully conducted investigations. In addition to the specific automatisms which are required for the four fundamental operations with integers, a number of other automatisms are required for the operations with fractions both common and decimal. At present we have only partial analysis of the examples in these fields, and for that reason it is not

possible to state what are the types of examples that are within the range of school work.

These abilities are specific habits or automatic responses. Their significant characteristics are the rate or speed of performance and the accuracy of the response. Thus, the measurement of arithmetical abilities in these fields involves determining only at what rate a pupil is able to do examples of the elemental types, and how accurate his answers are. This is accomplished by having him do examples of a given type for a specified time. From his test paper his speed and per cent of examples correct may be determined. These two quantities represent the measure of his ability to do this type of example.¹

A complete and detailed measurement would require that a test be provided for each type of example, but fortunately certain combinations can be made. An example in addition consisting of three columns of nine figures each includes the addition combinations, simple column addition, and carrying. Thus, if a pupil responds satisfactorily to examples of this type, we know that he possesses the ability to do the types of addition examples involved therein. On the other hand, if his response to this type of example is unsatisfactory, we do not know just what elemental ability he lacks. The use of a single test of this type to measure

¹ Strictly speaking the number of examples done and the per cent of examples correct is a measure of the pupil's performance rather than of his ability. A pupil's performance is affected by many factors such as his emotional status, physical condition, light, temperature, and the like. Or, it may be that a pupil does not try to do his best on a given test. A pupil's ability can only be inferred from his performance, but when conditions are properly controlled, such inference is reliable in all except a few cases. In order to avoid an awkward form of statement and because the practice is general, we shall speak of a score as a measure of a pupil's ability.

a group of arithmetical abilities has this very obvious limitation in diagnosing the conditions which exist, but it does provide a very satisfactory general survey.

Why we need to use arithmetical tests. The fundamental reason for measurement is to secure information which will be helpful in making instruction more effective. A general survey furnishes general information. Such information is useful in determining the general effectiveness of the instruction. The teacher, however, is primarily concerned with details of instruction and with individual pupils, and therefore must have detailed information in order to know how to adjust the instruction to the needs of the individual pupils. She needs to learn what types of examples her pupils can do with a satisfactory degree of facility, and what types they cannot do. She needs to learn what pupils possess standard ability and what pupils do not. A general test serves to locate the pupils who are not yet up to standard, but a more elaborate test must be used to reveal the exact nature of the shortcomings of the pupils.

Besides the abilities involved in performing the operations of arithmetic, there is another large group of abilities that function in determining what operations to perform in solving problems.¹ The analysis of this division of arithmetical abilities has not been carried as far as in the case of the operations of arithmetic. However, it appears that these abilities involve knowledge rather than specific habits.

¹ The word "problem" is used by some writers to designate both "examples" and "problems." In this book the word "example" will be used to designate exercises which explicitly call for certain arithmetical operations. The word "problem" will designate only those exercises which require the pupil to determine first what operations are to be performed.

II. STANDARDIZED TESTS FOR MEASURING ARITHMETICAL ABILITIES

1. The Courtis Standard Research Tests, Series B

The Standard Research Tests, Series B, or, as they are commonly called, the Courtis Arithmetic Tests, have probably been more widely used than any other instrument for measuring arithmetical abilities, and as a result we have better comparative standards for their use. The series consists of four tests, printed on four consecutive pages. They are suitable for a general survey of the abilities of pupils to perform the operations with integers.

Test No. 1. Addition

The twenty-four examples of this test have been constructed so that all have the same form, three columns of nine figures each. The following are samples of the examples. Time allowed, 8 minutes.

<u>927</u>	<u>297</u>	<u>136</u>	<u>486</u>	<u>384</u>	<u>176</u>
<u>379</u>	<u>925</u>	<u>340</u>	<u>765</u>	<u>477</u>	<u>783</u>
<u>756</u>	<u>473</u>	<u>988</u>	<u>524</u>	<u>881</u>	<u>697</u>
<u>837</u>	<u>983</u>	<u>386</u>	<u>140</u>	<u>266</u>	<u>200</u>
<u>924</u>	<u>315</u>	<u>353</u>	<u>812</u>	<u>679</u>	<u>366</u>
<u>110</u>	<u>661</u>	<u>904</u>	<u>466</u>	<u>241</u>	<u>851</u>
<u>854</u>	<u>794</u>	<u>547</u>	<u>355</u>	<u>796</u>	<u>535</u>
<u>965</u>	<u>177</u>	<u>192</u>	<u>834</u>	<u>850</u>	<u>323</u>
<u>344</u>	<u>124</u>	<u>439</u>	<u>567</u>	<u>733</u>	<u>229</u>

Test No. 2. Subtraction

This test consists of twenty-four examples, each involving the same number of subtractions. The following are samples. Time allowed, 4 minutes.

<u>107795491</u>	<u>75088824</u>	<u>91500053</u>	<u>87939983</u>
<u>77197029</u>	<u>57406394</u>	<u>19901563</u>	<u>72207316</u>

24 EDUCATIONAL TESTS AND MEASUREMENTS

Test No. 3. Multiplication

This test consists of twenty-four examples of this type.
Time allowed 6 minutes.

8246	3597	5739	2648
— 29	— 73	— 85	— 46
			— 92

Test No. 4. Division

This test consists of twenty-four examples of this type.
Time allowed, 8 minutes.

25) <u>6775</u>	94) <u>85352</u>	37) <u>9990</u>	86) <u>80066</u>
— 58	— 85	— 109	— 72
73) <u>58765</u>	49) <u>31409</u>	68) <u>43520</u>	52) <u>44252</u>
— 44	— 31	— 435	— 252

In giving the test the pupils are directed as follows: —

You will be given eight minutes to find the answers to as many of these addition examples as possible. Write the answers on this paper directly underneath the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples.

Marking the papers. In marking the test papers, which is done by the use of a printed answer card which is run along across the page, no credit is given for examples partly right nor for examples partly completed. A pupil's score is the number of examples attempted and the number right. This simple plan of marking the papers insures uniformity and accuracy.

Each of the examples of a test calls for the same number of operations under approximately the same conditions. This makes the examples of each test approximately equal in difficulty. Any example of the addition test, say the seventh,

is just as difficult as any other, say the second. Thus, the tests consist of twenty-four equal units, just as a yardstick consists of thirty-six equal units (inches). The measure of a pupil's ability is represented by the distance he advances along the scale in the given time, i.e., by the number of examples done and by the per cent of these examples which have been done correctly.

Since an example of one of these tests is defined as so many operations under certain conditions, it is possible to construct other tests equal in difficulty. Four forms have been constructed. This makes it possible to use a different form when the tests are given a second time.

2. The Cleveland-Survey Arithmetic Tests

These are a series of fifteen tests devised to analyze the arithmetical processes, and in this respect differ from the Courtis tests described above. The following statement, taken from the report of the director of the testing work of the Survey, will explain the nature of the tests devised:—

Spiral nature of the tests. The test, which was given to all of the A grades in the system, included a number of different forms of each of the fundamental operations. Thus, in addition, the first and simplest exercise of the test consisted in adding pairs of figures. Later in the series, addition appeared again, but in a more elaborate form. It was here required that a short column of figures be added. The third case of addition consisted in the adding of fractions of like denominators. The fourth case consisted in the addition of a longer column of figures. This differs from the short-column addition in the fact that a greater effort of attention is required in order to complete the addition. Addition of four-place figures which requires carrying forward from one column to the next, and addition of fractions of unlike denominators, constituted the final and most elaborate stages of the addition process. The purpose of introducing these various types of addition was to test

the ability of the different grades to perform increasingly elaborate operations. Similar spiral tests in subtraction, multiplication, and division were interwoven with the exercises in addition.

In the second place, the test was so presented that the rate of work in the different grades could be determined. . . . The test shows, therefore, both the complexity of the processes which a given grade can master, and also the number of examples of a given type that can be performed in a specified time.¹

These tests have also been used in the recent school surveys at Grand Rapids, Michigan, and St. Louis, Missouri.

Nature of the tests. The nature of the tests devised may be seen from the accompanying samples taken from each of the fifteen, on pages 27 and 28.

The time allowances for the several tests are as follows:—

Set A . . . 30 seconds	Set F . . . 1 minute	Set K . . . 2 minutes
Set B . . . 30 seconds	Set G . . . 1 minute	Set L . . . 3 minutes
Set C . . . 30 seconds	Set H . . . 30 seconds	Set M . . . 3 minutes
Set D . . . 30 seconds	Set I . . . 1 minute	Set N . . . 3 minutes
Set E . . . 30 seconds	Set J . . . 2 minutes	Set O . . . 3 minutes

As in the case of the Courtis tests the examples of each test are approximately equal in difficulty. Thus each test may be considered to consist of approximately equal units. In marking the test papers no credit is given for examples partly right nor for examples partly completed which insures uniformity and accuracy. A pupil's score is the number of examples attempted and the number right.

In considering the completeness of this series of tests it must be remembered that decimal fractions are omitted, and that two tests are certainly inadequate for the field of common fractions. These tests, however, furnish a means for securing more detailed measurements of the arithmetical abilities of pupils than are possible by using the Courtis Standard Research Tests, Series B.

¹ Judd, Chas. H. *Measuring the Work of the Public Schools*, pp. 95-96.

Set A. Addition

$$\begin{array}{r} \underline{1} & \underline{6} & \underline{9} & \underline{0} \\ \underline{2} & \underline{6} & \underline{5} & \underline{1} \end{array} \quad \begin{array}{r} \underline{4} & \underline{1} & \underline{7} & \underline{9} & \underline{3} \\ \underline{2} & \underline{3} & \underline{7} & \underline{6} & \underline{0} \end{array} \quad \begin{array}{r} \underline{2} & \underline{1} & \underline{3} & \underline{6} \\ \underline{4} & \underline{5} & \underline{8} & \underline{9} \end{array}$$

Set B. Subtraction

$$\begin{array}{r} \underline{9} & \underline{7} & \underline{11} & \underline{8} & \underline{12} & \underline{1} & \underline{9} \\ \underline{9} & \underline{3} & \underline{6} & \underline{1} & \underline{3} & \underline{0} & \underline{7} \end{array} \quad \begin{array}{r} \underline{13} & \underline{4} & \underline{12} \\ \underline{8} & \underline{3} & \underline{6} \end{array}$$

Set C. Multiplication

$$\begin{array}{r} \underline{2} & \underline{4} & \underline{9} & \underline{0} & \underline{5} & \underline{4} & \underline{2} \\ \underline{2} & \underline{7} & \underline{8} & \underline{2} & \underline{6} & \underline{1} & \underline{9} \end{array} \quad \begin{array}{r} \underline{7} & \underline{4} & \underline{9} \\ \underline{6} & \underline{0} & \underline{5} \end{array}$$

Set D. Division

$$\begin{array}{r} 3) \underline{9} \\ 4) \underline{32} \\ 6) \underline{36} \\ 2) \underline{0} \\ 7) \underline{28} \\ 9) \underline{9} \\ 3) \underline{21} \end{array}$$

Set E. Addition

$$\begin{array}{r} \underline{5} & \underline{2} & \underline{9} & \underline{2} & \underline{6} & \underline{1} & \underline{4} & \underline{9} \\ \underline{2} & \underline{8} & \underline{8} & \underline{8} & \underline{3} & \underline{4} & \underline{6} & \underline{7} \\ \underline{2} & \underline{8} & \underline{0} & \underline{5} & \underline{4} & \underline{2} & \underline{5} & \underline{1} \\ \underline{0} & \underline{5} & \underline{7} & \underline{0} & \underline{8} & \underline{5} & \underline{3} & \underline{5} \\ \underline{4} & \underline{1} & \underline{6} & \underline{6} & \underline{8} & \underline{4} & \underline{4} & \underline{3} \end{array}$$

Set F. Subtraction

$$\begin{array}{r} 616 \\ 456 \\ \hline \end{array} \quad \begin{array}{r} 1248 \\ 709 \\ \hline \end{array} \quad \begin{array}{r} 1365 \\ 618 \\ \hline \end{array} \quad \begin{array}{r} 1092 \\ 472 \\ \hline \end{array} \quad \begin{array}{r} 716 \\ 344 \\ \hline \end{array}$$

Set G. Multiplication

$$\begin{array}{r} 2345 \\ 2 \\ \hline \end{array} \quad \begin{array}{r} 9735 \\ 5 \\ \hline \end{array} \quad \begin{array}{r} 8642 \\ 9 \\ \hline \end{array} \quad \begin{array}{r} 6789 \\ 2 \\ \hline \end{array} \quad \begin{array}{r} 2345 \\ 6 \\ \hline \end{array}$$

Set H. Fractions

$$\frac{3}{5} + \frac{1}{5} = \quad \frac{6}{9} - \frac{4}{9} = \quad \frac{4}{9} + \frac{1}{9} = \quad \frac{8}{9} - \frac{7}{9} =$$

28 EDUCATIONAL TESTS AND MEASUREMENTS

Set I. Division

4)55424

7)65982

2)58748

5)41780

Set J. Addition

<u>7</u>	<u>9</u>	<u>4</u>	<u>7</u>	<u>2</u>	<u>9</u>	<u>6</u>	<u>7</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>4</u>	<u>3</u>	<u>2</u>
<u>5</u>	<u>2</u>	<u>5</u>	<u>1</u>	<u>9</u>	<u>6</u>	<u>9</u>	<u>1</u>	<u>8</u>	<u>0</u>	<u>5</u>	<u>3</u>	<u>1</u>	<u>1</u>
<u>4</u>	<u>4</u>	<u>8</u>	<u>9</u>	<u>4</u>	<u>2</u>	<u>6</u>	<u>5</u>	<u>5</u>	<u>7</u>	<u>3</u>	<u>7</u>	<u>7</u>	<u>6</u>
<u>2</u>	<u>8</u>	<u>1</u>	<u>4</u>	<u>8</u>	<u>4</u>	<u>7</u>	<u>1</u>	<u>4</u>	<u>1</u>	<u>4</u>	<u>7</u>	<u>6</u>	<u>6</u>
<u>6</u>	<u>2</u>	<u>4</u>	<u>3</u>	<u>5</u>	<u>7</u>	<u>0</u>	<u>4</u>	<u>1</u>	<u>8</u>	<u>6</u>	<u>0</u>	<u>9</u>	<u>1</u>
<u>0</u>	<u>7</u>	<u>8</u>	<u>2</u>	<u>1</u>	<u>1</u>	<u>4</u>	<u>6</u>	<u>8</u>	<u>5</u>	<u>2</u>	<u>2</u>	<u>6</u>	<u>8</u>
<u>5</u>	<u>5</u>	<u>5</u>	<u>8</u>	<u>5</u>	<u>3</u>	<u>3</u>	<u>5</u>	<u>2</u>	<u>1</u>	<u>3</u>	<u>9</u>	<u>3</u>	<u>6</u>
<u>1</u>	<u>3</u>	<u>1</u>	<u>5</u>	<u>2</u>	<u>9</u>	<u>7</u>	<u>3</u>	<u>1</u>	<u>3</u>	<u>9</u>	<u>5</u>	<u>4</u>	<u>9</u>
<u>8</u>	<u>6</u>	<u>3</u>	<u>2</u>	<u>4</u>	<u>2</u>	<u>1</u>	<u>3</u>	<u>3</u>	<u>7</u>	<u>2</u>	<u>6</u>	<u>5</u>	<u>7</u>
<u>3</u>	<u>1</u>	<u>9</u>	<u>7</u>	<u>3</u>	<u>3</u>	<u>6</u>	<u>7</u>	<u>9</u>	<u>4</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
<u>2</u>	<u>4</u>	<u>6</u>	<u>7</u>	<u>6</u>	<u>8</u>	<u>0</u>	<u>6</u>	<u>8</u>	<u>9</u>	<u>8</u>	<u>4</u>	<u>2</u>	<u>2</u>
<u>9</u>	<u>8</u>	<u>3</u>	<u>1</u>	<u>7</u>	<u>5</u>	<u>6</u>	<u>1</u>	<u>4</u>	<u>4</u>	<u>5</u>	<u>8</u>	<u>9</u>	<u>2</u>
<u>9</u>	<u>8</u>	<u>5</u>	<u>9</u>	<u>6</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>5</u>	<u>4</u>	<u>6</u>	<u>8</u>	<u>9</u>	<u>4</u>

Set K. Division

21)441

32)672

23)483

51)1173

Set L. Multiplication

8246
29

3597
73

5739
85

2648
46

9537
92

Set M. Addition

<u>7493</u>	<u>8937</u>	<u>8625</u>	<u>2123</u>	<u>5142</u>	<u>3691</u>
<u>9016</u>	<u>6345</u>	<u>4091</u>	<u>1679</u>	<u>0376</u>	<u>4526</u>
<u>6487</u>	<u>2783</u>	<u>3844</u>	<u>5555</u>	<u>4955</u>	<u>7479</u>
<u>7591</u>	<u>4883</u>	<u>8697</u>	<u>6331</u>	<u>9314</u>	<u>2087</u>
<u>6166</u>	<u>1341</u>	<u>7314</u>	<u>6808</u>	<u>5507</u>	<u>8165</u>

Set N. Division

67)32763

48)28464

97)36084

59)29382

Set O. Fractions

$$\frac{11}{15} + \frac{1}{6} =$$

$$\frac{9}{14} - \frac{1}{4} =$$

$$\frac{3}{4} \times \frac{5}{6} =$$

$$\frac{20}{21} \div \frac{1}{6} =$$

3. *The Woody Arithmetic Scales*

Measurement by means of a scale. Woody has recently devised a set of four scales, one for each of the four fundamental operations. He states that his "fundamental idea was to derive a series of scales which would indicate the type of problems (examples) and the difficulty of the problems (examples) that a class can solve correctly."¹ The addition scale of Series A is reproduced on page 31. The character of the examples and their arrangement are the same in the other scales.² Twenty minutes are allowed for each scale of Series A, and ten minutes for each scale of Series B. This time allowance is sufficient for most pupils in grades above the fourth to complete all of the examples.

The difficulty of each example has been determined and the examples of each scale are arranged in order of increasing difficulty. The author makes only this statement concerning the selection of the examples. "Each of the scales is composed of as great a variety of problems (examples) as the fundamental operations can well permit," and only those examples "were chosen which were solved by a gradually increasing percentage of the pupils as one proceeded from the lower to the higher grades."

It should be noted that a "scale" differs fundamentally from the tests described above. The examples of a scale are not equally difficult. In the case of Woody's scales the score of a pupil is a statement of the particular examples which he has done correctly. The score of a class is the degree of

¹ Woody, Clifford. *Measurements of Some Achievements in Arithmetic*, p. 1. (Teachers College Contribution to Education, no. 80. 1916.)

² There are two series of four scales each. Series B differs from Series A in that it consists of only certain examples taken from Series A.

difficulty of the example which was done correctly by just 50 per cent of the class. Ability as measured by these scales means simply that certain types of examples can be done correctly and that certain other types cannot be done correctly. The speed at which the examples can be done is not included in the meaning of ability. Thus "ability" when used in connection with the Woody Arithmetic Scales cannot have the same meaning as is attached to the word when used in connection with the Courtis Standard Research Tests, Series B, or other tests of the same type.

On the basis of such analyses of arithmetical abilities as have been made, it is clear that all types of examples have not been included. It also appears from Woody's own statement that those which were chosen were selected not on the basis of their arithmetical significance but on the basis of the consistency of pupils' reactions.

Woody's method of selecting examples may be called *statistical* as opposed to the *analytical* method employed by Courtis and other makers of tests. The statistical method neglects the subject-matter field in which the test is being constructed and assumes that an example is suitable for use in a test simply because it is done correctly by a gradually increasing per cent of pupils as one proceeds from grade to grade. It rejects as unfit those examples which do not have this characteristic.¹ On the other hand, the analytical method involves a careful analysis of the field of subject-matter in which the test is being constructed to determine the fundamental types of examples which exist. This method

¹ From what we know about the curve of learning it is doubtful whether this basis can be justified. Ability to do does not increase gradually from grade to grade.

ADDITION SCALE

31

Name.....

When is your next birthday? How old will you be?

Are you a boy or girl?..... In what grade are you?.....

$$\begin{array}{r}
 \begin{array}{ccccccccc}
 (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) \\
 \underline{2} & \underline{2} & 17 & 53 & 72 & 60 & 3+1 = & 2+5+1 = & 20 \\
 - & - & - & - & - & - & & & \\
 \underline{3} & \underline{4} & \underline{2} & \underline{45} & \underline{26} & \underline{37} & & & \\
 - & - & - & - & - & - & & & \\
 \underline{1} & \underline{3} & & & & & & & \\
 - & - & & & & & & & \\
 \end{array}
 \end{array}$$

(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	25
21	32	43	23	25 + 42 =	100	9	199	2563	
33	59	1	25		33	24	194	1387	
35	17	2	16		45	12	295	4954	
—	—	13	—		201	15	156	2065	
					46	19			

(19)	(20)	(21)	(22)	(23)	(24)	(25)
\$.75	\$12.50	\$8.00	547	$\frac{1}{8} + \frac{1}{8} =$	4.0125	$\frac{8}{8} + \frac{5}{8} + \frac{7}{8} + \frac{1}{8} =$
1.25	16.75	5.75	197		1.5907	
.49	15.75	2.33	685		4.10	
		4.16	678		8.673	
		.94	456			
		6.32	393			
			525			
			240			
			152			

$$\begin{array}{r}
 (26) \quad 12\frac{1}{2} \quad (27) \quad \frac{1}{8} + \frac{1}{4} + \frac{1}{2} = \\
 62\frac{1}{2} \\
 37\frac{1}{2} \\
 \hline
 \end{array}
 \quad
 \begin{array}{r}
 (28) \quad \frac{8}{4} + \frac{1}{4} = \\
 2\frac{1}{4} \\
 5\frac{1}{4} \\
 \hline
 \end{array}
 \quad
 \begin{array}{r}
 (29) \quad 4\frac{3}{4} \\
 2\frac{1}{2} \\
 3\frac{3}{4} \\
 \hline
 \end{array}
 \quad
 \begin{array}{r}
 (30) \quad 113.46 \\
 6\frac{2}{3} \\
 19.9 \\
 \hline
 \end{array}
 \quad
 \begin{array}{r}
 (31) \quad .0086 \\
 18.253 \\
 6.04
 \end{array}
 \quad
 \begin{array}{r}
 (32) \quad \frac{3}{4} + \frac{1}{2} + \frac{1}{4} =
 \end{array}$$

(33)				
.49	(34)	(35)	(36)	(37)
.28	$\frac{1}{8} + \frac{3}{8} =$	2 ft. 6 in.	2 yr. 5 mo.	$16\frac{1}{3}$
.63		3 ft. 5 in.	3 yr. 6 mo.	$12\frac{1}{8}$
.95		<u>4 ft. 9 in.</u>	4 yr. 9 mo.	$21\frac{1}{2}$
1.69			5 yr. 2 mo.	$32\frac{3}{4}$
.22			6 yr. 7 mo.	

is well illustrated in the derivation of the Addition of Fractions Tests by Ballou. These tests are described on page 34.

Woody makes the following statement with reference to the uses of the scales: —

Perhaps the most valuable use of the scales lies in the diagnosing power of the class mistakes. The writer was convinced during the process of scoring these test papers, nearly 20,000 in all, that the mistakes of a class tend to be grouped around some central tendency. The great variety of the problems in these scales, and the fact that the problems in each of the various operations proceed from the simplest to the more difficult problems, aid greatly in the location of the weaknesses of the class. If a large number in a class fail to invert the divisor in the problems in division of fractions, or if a large number in a class fail to locate the decimal point properly in the problems in multiplication of decimal fractions, a teacher should know immediately that these classes need more practice in these particular processes. In a like manner, by locating the particular types of problems missed, one should be able to direct the work of a class more intelligently.

To obtain a diagnosis of a class it is necessary to tabulate the results for each example. The score sheet for a typical class is given in Table II. The examples not listed in the tabulation were not done incorrectly by more than two pupils. An example not attempted is indicated by a dash. An example done incorrectly is indicated by 1. By examining the per cent of examples right at the bottom of the table one learns the types of examples on which this class needs instruction.

In view of the manner in which the examples for the scales were chosen, it seems reasonable that certain limitations should be placed upon Woody's claim for the diagnosing power of these scales. It may also be questioned whether one example is sufficient to test adequately the

TABLE II

SHOWING THE TABULATION OF THE SCORES MADE BY THE INDIVIDUAL PUPILS OF A SIXTH-GRADE CLASS UPON THE WOODY ADDITION SCALE, SERIES A

<i>Pupil</i>	<i>Examples</i>																		
	18	20	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1.....					1				1	1	1		1						
2.....					1				-	1	1								
3.....		1	1	1	1			1	1		1		1		1		1	-	
4.....		1		1											1				
5.....												1		1			1	1	
6.....		1										1		1			1		
7.....				1		1				1	1	1				1	1		
8.....			1					1				1		1	1	1	1		1
9.....		1			1	1	1	1	1		1		1		1		1	1	
10.....				1	1	1	1	1	1		1	1	1		1		1	1	1
11.....																			
12.....																1		1	1
13.....			1			1	1	1	1	1	1	1	1	1	1		1	1	
14.....		1	1		1		1	1	1	1	1	1	1	1	-	-	-	-	
15.....		1																	1
16.....								1			1					1		1	-
17.....	1								1		1					1		1	1
18.....			1			1	1	1	1	1	1	1	1	1	1		1	1	1
19.....								1											-
20.....			1					1			1	1	1		1	1			-
21.....									1		1	1		1	1			1	
22.....	1	1	1	1		1						1			1	1	1	1	1
23.....	1		1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
24.....			1				1	1		1	1	1	1			1		1	
25.....			1				1	1	1	1	1	1	1	1		1	1	1	1
26.....		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27.....									1					1					
28.....		1		1				1			1	1	1	1	1		1	1	1
Per cent right	88	88	57	68	75	65	65	43	72	57	36	46	53	61	34	82	67	37	33

ability of a class to do examples of that type.¹ For instance, the addition combinations are represented by only these two, $2 + 3$ and $3 + 1$. Finally, it should be remembered that of the characteristics of specific abilities, accuracy only, is measured. The time allowed is sufficient for most pupils to complete the test.

4. Research Tests in Arithmetic; Addition of Fractions

These tests, devised by F. W. Ballou, Director of the Bureau of Educational Investigations and Measurement, of the Boston Public Schools, furnish a good illustration of tests based upon a careful analysis of abilities. The analysis of the addition of two fractions revealed 14 types of examples,² which arise out of reducing the fractions to a common denominator and reducing the answer to the lowest form. This analysis was corroborated by a preliminary testing of pupils. It was found that pupils could do certain types of examples and fail on others, showing thereby that ability to do examples of one type did not function efficiently in doing examples of other types. It is obvious that the addition of three or more fractions involves a number of other types of examples. It is also obvious that subtraction, multiplication, and division each involve a number of types of examples.

In order that both speed and accuracy may be measured, a separate test is needed for each type of example. Certain types of examples are included in others. The example

¹ Recently the writer has collected data which indicates that the diagnosis secured by the use of Woody's Scales is defective in this respect.

² *Arithmetic. Determining the Achievements of Pupils in the Addition of Fractions.* (School Document no. 3. 1916. Boston Public Schools.)

$\frac{1}{2} + \frac{3}{10}$ includes examples of the types $\frac{1}{2} + \frac{3}{8}$ and $\frac{5}{9} + \frac{1}{9}$. Recognizing this fact, Ballou has constructed a series of six tests to measure in detail the ability of pupils to add two fractions. Each test is illustrated by two examples. The time allowance for each test is two minutes.

ADDITION OF FRACTIONS

Test 1

(1) $\frac{1}{4}$	(2) $\frac{3}{14}$
$\frac{1}{4}$	$\frac{1}{14}$
$\frac{1}{4}$	$\frac{1}{14}$
$\underline{\quad}$	$\underline{\quad}$

Test 2

(1) $\frac{1}{3}$	(2) $\frac{2}{7}$
$\frac{1}{3}$	$\frac{2}{7}$
$\frac{1}{3}$	$\frac{2}{7}$
$\underline{\quad}$	$\underline{\quad}$

Test 3

(1) $\frac{3}{5}$	(2) $\frac{5}{6}$
$\frac{3}{5}$	$\frac{5}{6}$
$\frac{11}{15}$	$\frac{1}{2}$
$\underline{\quad}$	$\underline{\quad}$

Test 4

(1) $\frac{1}{7}$	(2) $\frac{7}{9}$
$\frac{1}{7}$	$\frac{7}{9}$
$\frac{9}{10}$	$\frac{1}{4}$
$\underline{\quad}$	$\underline{\quad}$

Test 5

(1) $\frac{1}{10}$	(2) $\frac{4}{9}$
$\frac{1}{10}$	$\frac{4}{9}$
$\frac{1}{6}$	$\frac{5}{12}$
$\underline{\quad}$	$\underline{\quad}$

Test 6

(1) $\frac{1}{6}$	(2) $\frac{5}{6}$
$\frac{1}{6}$	$\frac{5}{6}$
$\frac{9}{10}$	$\frac{3}{8}$
$\underline{\quad}$	$\underline{\quad}$

5. *The Stone Reasoning Test*

Several tests have been devised to measure the abilities of pupils to solve problems involving reasoning but none of them have proven very satisfactory. Some years ago Stone¹

¹ Stone, C. W. *Arithmetical Abilities and Some Factors Determining Them.* (Teachers College Contributions to Education, no. 19. 1908.) See also Stone, C. W., *Standardized Reasoning Tests in Arithmetic and How to Utilize Them.* (Teachers College Contributions to Education, no. 83. 1916.)

worked out a reasoning test which has been used in several cities, and in a number of city school surveys, so that we have rather definite standards as to what may be expected from its use.

THE STONE REASONING TEST

School..... Grade..... Name of Pupil.....

<i>Prob- lem value</i>	<i>Problems</i>
	Solve as many of the following problems as you have time for; work them in order as numbered:
1.0	1. If you buy 2 tablets at 7 cents each and a book for 65 cents, how much change should you receive from a two-dollar bill?
1.0	2. John sold 4 Saturday Evening Posts at 5 cents each. He kept $\frac{1}{2}$ the money and with the other $\frac{1}{2}$ he bought Sunday papers at 2 cents each. How many did he buy?
1.0	3. If James had 4 times as much money as George, he would have \$16. How much money has George?
1.0	4. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?
1.0	5. The uniforms for a baseball nine cost \$2.50 each. The shoes cost \$2 a pair. What was the total cost of uniforms and shoes for the nine?
1.4	6. In the schools of a certain city there are 2200 pupils; $\frac{1}{2}$ are in the primary grades, $\frac{1}{4}$ in the grammar grades, $\frac{1}{8}$ in the High School and the rest in the night school. How many pupils are there in the night school?
1.2	7. If $3\frac{1}{2}$ tons of coal cost \$21, what will $5\frac{1}{2}$ tons cost?
1.6	8. A news dealer bought some magazines for \$1. He sold them for \$1.20, gaining 5 cents on each magazine. How many magazines were there?
2.0	9. A girl spent $\frac{1}{8}$ of her money for car fare, and three times as much for clothes. Half of what she had left was 80 cents. How much money did she have at first?
2.0	10. Two girls receive \$2.10 for making buttonholes. One makes 42, the other 28. How shall they divide the money?
2.0	11. Mr. Brown paid one third of the cost of a building; Mr. Johnson received \$500 more annual rent than Mr. Brown. How much did each receive?
2.0	12. A freight train left Albany for New York at 6 o'clock. An express left on the same track at 8 o'clock. It went at the rate of 40 miles an hour. At what time of day will it overtake the freight train if the freight train stops after it has gone 56 miles?

The time allowance for the test is fifteen minutes. Stone's plan for marking the test papers allows credit for examples partly right and for examples which were not finished. The problem values have been determined upon the basis of difficulty. It should be noted that this plan for marking the test papers is not as simple as that employed for marking the test papers on the operations of arithmetic.

6. *Other Reasoning Tests*

Courtis included two reasoning tests¹ in his Series A. Starch has devised a test which is called Arithmetical Scale A.² This scale included a number of the problems used by Stone, Courtis, and Thorndike. They have been evaluated upon the basis of difficulty and arranged in order of increasing difficulty. The pupils are allowed as much time as they need and a pupil's score is the value of the most difficult problem done correctly.

The problem field of arithmetic has not yet been analyzed and we do not know what the fundamental types of problems are. A partial analysis³ indicates that there are many types of problems of which probably a relatively few are fundamental. Until the fundamental types of problems are determined by analysis it will not be possible to devise a test which will be as satisfactory as the tests which we now have

¹ Courtis, S. A. *Manual of Instructions for Giving and Scoring the Courtis Standard Tests in the Three R's*. (Detroit, 1914.) These tests are no longer published.

² Starch, Daniel. "A Scale for Measuring Ability in Arithmetic," in *Journal of Educational Psychology*, vol. 7, pp. 213-22.

³ Monroe, Walter S. "A Preliminary Report of an Investigation of the Economy of Time in Arithmetic"; *Sixteenth Yearbook of the National Society for the Study of Education*, part I.

for the operations of arithmetic. Although it is probably wise to use the reasoning tests which are now available their limitations should be kept in mind.

III. STANDARD SCORES

The degree of ability which a pupil of a given grade should possess is called a standard. Standards are necessary to give meaning to the scores which pupils make. In most cases the standards are median¹ or average scores and thus represent merely the consensus of present practice. Such standards are open to the criticism that we cannot be certain that our present practice is satisfactory, but it seems probable that standards derived in this way will not be changed materially in the near future provided they are based upon a sufficient number of cases. The topic of standards and the use of them is discussed at length in Chapter IX and the reader may profitably study it in connection with the topic of standard scores in this and the following chapters.

1. Courtis Standard Research Tests, Series B.

Standard Median Scores. In Table III there are given three standard scores: (1) general median scores based upon distributions of "many thousands of individual scores in tests given in May or June, 1915-16. The distribution for each grade was made up of approximately equal numbers of classes from large-city schools and from small-city and country schools"; (2) the standards proposed by Courtis after three years' use of these tests;

¹ See page 242 for definition of "median."

(3) Boston standard median scores after the tests had been used for three years.

With reference to the standards which he has proposed Courtis says: —

The speeds set as standard are approximately the average speeds at which the children of the different grades have been found to work when tested at the end of the year, when for any one grade a random selection of five thousand scores from children in schools of all types and kinds are used as a basis of judgment.

Standard accuracy is perfect work, one hundred per cent. This is a tentative standard only, as there is available very little information in regard to the factors that determine accuracy and the effects of more efficient training.

At present in addition and multiplication it is only very exceptional work in which the median rises above eighty per cent accuracy, while in subtraction and division the limiting level is ninety per cent.

Standard speeds are not likely to change greatly. Standard accuracy is surely destined to approach much more nearly one hundred per cent than present work would indicate.

Standard scores are not only goals to be reached; they are limits not to be exceeded. It seems as foolish to overtrain a child as it is to undertrain him. All direct drill work should, in the judgment of the writer, be discontinued once the individual has reached standard levels. If his abilities develop further through incidental training, well and good, but the superintendent who, by repeated raising of standards, forces teachers and pupils to spend each year a larger percentage of time and effort upon the mere mechanical skills, makes as serious a mistake as the superintendent who is too lax in his standards.¹

Comparisons with these standards or any others are valid only when the tests have been given under standard conditions. Slight changes in the method of giving the tests may affect the scores as much as the difference in the standards from one grade to another.

¹ Courtis, S. A. *Third, Fourth, and Fifth Annual Accounting, 1913-16* (Department of Coöoperative Research, Detroit), p. 49.

TABLE III. STANDARD MEDIAN SCORES, STANDARD RESEARCH TESTS, SERIES B

Grade		Addition		Subtraction		Multiplication		Division	
		Speed	Accuracy	Speed	Accuracy	Speed	Accuracy	Speed	Accuracy
IV . . .	General	7.4	64	7.4	80	6.2	67	4.6	57
	Courtis	6	100	7	100	6	100	4	100
	Boston	8	70	7	80	6	60	4	60
V . . .	General	8.6	70	9.0	83	7.5	75	6.1	77
	Courtis	8	100	9	100	8	100	6	100
	Boston	9	70	9	80	7	70	6	70
VI . . .	General	9.8	73	10.3	85	9.1	78	8.2	87
	Courtis	10	100	11	100	9	100	8	100
	Boston	10	70	10	90	9	80	8	80
VII . . .	General	10.9	75	11.6	86	10.2	80	9.6	90
	Courtis	11	100	12	100	10	100	10	100
	Boston	11	80	11	90	10	80	10	90
VIII . . .	General	11.6	76	12.9	87	11.5	81	10.7	91
	Courtis	12	100	13	100	11	100	11	100
	Boston	12	80	12	90	11	80	11	90

Speed is the number of examples done in the time allowed.

Accuracy is the per cent of examples correct.

"General" medians were determined by Courtis on the basis of the 1916 tabulations and summaries of tabulations of other years. Courtis, S. A. *Third, Fourth, and Fifth Annual Accountings, 1913-16*. (Department of Coöperative Research, Detroit.)

The Boston standards were established after using the tests for three years. Ballou, F. W. *Arithmetic, the Courtis Standard Tests in Boston, 1912-15*. (Bulletin No. 10 of the Department of Educational Investigation and Measurement.)

2. The Cleveland-Survey Tests

Cleveland and Grand Rapids scores. The tests which we have designated as the Cleveland-Survey tests have also been used in the school survey made at Grand Rapids, Michigan. To show what results may be expected from the use of these tests we give, in Table IV, the median scores obtained

in Cleveland and Grand Rapids. In all cases the medians are for the lower half of the grades. Upper numbers, medians for Cleveland, Ohio. Lower numbers, medians for Grand Rapids, Michigan.

TABLE IV. SHOWING THE MEDIAN SCORES FOR THE CLEVELAND-SURVEY ARITHMETIC TESTS

<i>Test</i>	<i>Grade</i>					
	<i>3b</i>	<i>4b</i>	<i>5b</i>	<i>6b</i>	<i>7b</i>	<i>8b</i>
A	13.4 11.8	17.8 13.6	22.2 20.3	24.8 22.8	26.7 26.5	27.5 29.5
B	9.3 6.3	13.4 9.1	17.2 14.7	19.8 16.8	21.5 21.3	26.0 22.8
C	6.5	12.0 7.1	15.5 13.7	16.6 15.5	17.7 17.7	19.0 19.3
D	6.3	12.4 6.9	15.7 12.5	18.5 15.5	20.8 18.4	22.5 20.5
E	4.3	5.3 4.1	6.3 5.2	6.8 6.0	7.5 7.2	7.8 7.8
F	2.0	4.9 2.8	6.7 6.0	7.5 7.1	8.6 9.3	10.1 10.3
G	2.0	3.9 2.2	5.2 4.5	5.5 5.3	5.9 6.1	6.6 6.7
H	0.0	0.0	5.0	5.5 6.2	7.7 9.0	8.5 8.6
I	0.6	1.1 0.7	2.0 1.3	3.1 2.3	4.0 3.8	4.7 4.0
J	1.9	3.2	4.0 3.4	4.1 4.1	4.9 5.4	5.7 5.7
K	0.0	4.0	6.8 3.0	8.5 5.4	10.1 7.5	12.5 9.7
L	0.0	1.7	2.5 2.3	2.8 3.3	3.2 4.3	3.9 4.9
M	1.4	2.5	3.2 3.0	3.8 4.3	4.4 4.9	5.1 5.7
N	0.0	0.8	1.3 0.7	1.7 1.1	2.0 1.7	2.6 2.0
O	0.0	0.0	0.0	3.1 3.5	4.1 3.9	5.5 5.5

3. The Woody Arithmetic Scales

No standards are given for the Woody scales for the reason that due to their recent development, only tentative standards are as yet available.

4. The Addition of Fractions Tests

The following tentative standards have been worked out in Boston from the use of these tests:—

TABLE V. BOSTON MEDIANs: ADDITION OF FRACTIONs

Grade	Pupils Tested	Test 1		Test 2		Test 3		Test 4		Test 5		Test 6	
		Speed Medians	Accuracy Medians										
VI...	1205	10.7	79.6	7.7	65.6	5.5	41.9	4.0	69.5	4.6	51.0	4.4	48.6
VII..	1243	16.5	86.6	10.1	72.9	7.3	46.1	5.3	69.2	6.3	54.9	5.7	48.1
VIII.	1130	20.7	88.2	11.6	74.4	8.4	47.4	6.0	67.8	6.9	52.4	6.4	46.5

5. The Stone Reasoning Test

While this test has been used in many cities, in few have all of the upper grades been tested and the records been kept separately by grades. Stone tested in twenty-six different cities, but used only the 6A grade. The test has been applied by others in comparison, but also using only the 6th grade. The sixth-grade scores for the twenty-six cities tested by Stone give the following results:—

Lowest 3.56

Middle 5.50

Highest 9.14

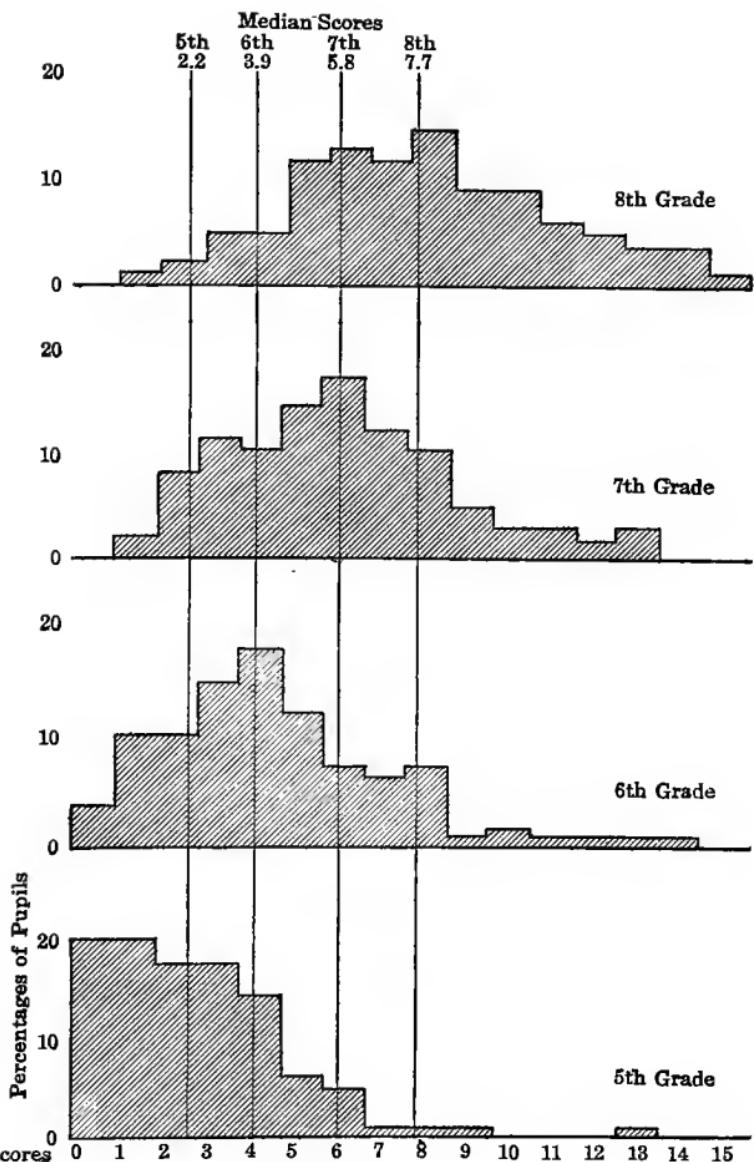


FIG. 2. DISTRIBUTION OF SCORES WITH THE STONE REASONING TESTS IN BUTTE

Representing the percentage of children making the given scores in reasoning problems. For example, 19 per cent of the fifth-grade children made a score of 0; 19 per cent made a score of 1, etc. The lines representing the median scores for each grade tell about how many in each grade surpass the median scores for the grades above, and how many fall below the median scores for the grades below.

In three cities where school surveys have been made recently the scores were taken separately by grades. The results in these three cities are given in Table VI. The wide distribution of scores for Butte is shown in Fig. 2, given on the opposite page.

TABLE VI. MEDIAN SCORES USING THE STONE REASONING TESTS

<i>Grade</i>	<i>Butte, Montana*</i>	<i>Bridgeport, Conn.†</i>	<i>Salt Lake City, Utah‡</i>
V.....	2.2	6.1	3.7
VI.....	3.9	5.2	6.4
VII.....	5.8	6.8	8.6
VIII.....	7.7	4.5	10.5

* *Report of a Survey of the School System of Butte, Montana*, p. 88. (1914.)

† *Report of the Examination of the School System of Bridgeport, Conn.*, p. 102. (1913.)

‡ *Report of a Survey of the School System of Salt Lake City, Utah*, p. 183. (1915.)

Other evidence, drawn from the survey work in these cities, would indicate that the children of Butte were low in ability to reason, the children of Salt Lake City high, and the children of Bridgeport quite uneven.

Recently Stone has issued the following standards:—

That 80 per cent or more of 5th grade pupils reach or exceed a score of 5.5 with at least 75 per cent accuracy; that 80 per cent or more of 6th grade pupils reach or exceed a score of 6.5 with at least 80 per cent accuracy; that 80 per cent or more of 7th grade pupils reach or exceed a score of 7.5 with at least 85 per cent accuracy; that 80 per cent or more of 8th grade pupils reach or exceed a score of 8.75 with at least 90 per cent accuracy.¹

The accuracy of individual scores. In considering the accuracy or preciseness of the measures obtained by the use of these tests, it should first be noted that in most of these

¹ Stone, C. W. *Standardized Reasoning Tests in Arithmetic and How to Utilize Them*. (*Teachers College Contributions to Education*, no. 83, 1916.)

tests and the manner of giving them the sources of error mentioned on pages 8-13 are eliminated. The plan of marking all examples as either right or wrong insures uniformity and accuracy in marking the papers. The rate at which the pupil works is measured as well as the quality of his work. The exercises are either equal in value or their difficulty has been expressed in terms of a common unit. By providing each pupil with a printed list of the examples, copying the example either from dictation or from the board is eliminated. It has been shown that pupils do not copy accurately nor do they copy with equal speed. The elimination of copying the examples eliminates a probable source of error.

There are, however, other factors to be considered. A pupil's performance or actual achievement from which his ability in a given test is inferred depends upon his physical, mental, and emotional condition. These change from day to day, and from hour to hour and cause marked variation in successive scores of some pupils. This variation is much greater in some pupils than in others. At any particular time a small per cent of the pupils will be found upon a plane higher than their normal ability, while other pupils will be found at the low ebb of their ability. This fact causes some of the measures to be unreliable in the sense that they are not true indices of the average or normal abilities of certain pupils. However, this happens in only a relatively small per cent of the cases when care is exercised in giving the tests to secure standard conditions, and the number of these cases can be materially reduced by giving the tests a second time and taking the average of the two sets of scores.

Courtis¹ states that "about one child in ten will have a markedly unreliable score," and "for two thirds of the children the differences will be relatively small."

Gain or loss in repeating tests. To test the accuracy of the individual scores, and the estimate of Courtis, the writer had the four Courtis tests of Series B given to the pupils in one school a second time. In order that the pupils might not do the same examples, they were asked, on the occasion of the second trial, to begin with the last example and work forward. The results are given in Table VII. For addition, the table is read as follows: One pupil did eight fewer examples the second trial, three did four fewer, four did three fewer, five did two fewer, etc. The addition scores average .15 of an example less on the second trial than they did on the first. Eighty-two per cent of the scores differed by one example, or agreed. Although this table includes too few cases to warrant any final conclusions, it is interesting to note that the facts are in accord with the statement of Courtis.

IV. HOW TO HANDLE WHAT THE TESTS REVEAL

Scientific management. During the past few years scientific management has been applied to many forms of human endeavor with results which have been nothing short of marvelous. For example, bricklaying has been practiced by intelligent artisans for centuries, and one might suppose that in the course of that length of time a highly efficient system of laying brick would have been evolved on the basis

¹ Courtis, S. A. "Single Measurements with Standard Tests"; in *Elementary School Teacher*, vol. 13, pp. 326-45, 486-504.

TABLE VII

SHOWING THE GAIN OF THE SECOND SCORE OVER THE FIRST
 (Repetition of Courtis tests, Series B, in one school.)

Gain																Per cent of scores differing by one ex- ample or the same		
	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	Total		
Addition....	-	-	1	-	-	3	4	5	26	37	19	3	1	1	1	100	-15	82
Subtraction...	-	1	-	-	-	1	12	8	22	21	19	10	4	1	1	99	-10	62
Multiplication	-	-	-	-	-	-	-	2	2	17	41	24	11	2	-	99	.16	82
Division.....	1	-	-	-	-	-	1	11	17	22	22	18	4	1	1	98	-28	62
Total.....	1	1	1	-	-	-	5	29	32	87	121	80	28	8	3	396	-	-
Per cent.....	.3	.3	.3	-	-	-	1.3	7.3	8.1	22.0	30.6	20.2	7.0	2.0	.8	-	-	72.8

of accidental successes and imitation, if in no other way. It appears, however, that the method followed has remained practically unchanged for centuries until recently, when the principles of scientific management were applied to the process. A scientific analysis of the process revealed that eighteen motions were made in laying each brick, while only five motions were needed when the material was properly arranged.

Another striking illustration is given by Frederick W. Taylor in his book, *The Principles of Scientific Management*. Some years ago a large quantity of pig iron was being loaded on flat cars at the Bethlehem steel plant. Pig iron is cast in blocks, each of which weighs ninety-two pounds. The method of loading was for the workman to pick up a pig, walk up an incline, deposit it upon the car, walk down, and repeat. The average amount of pig iron loaded per man per day was twelve and one half tons. It was very crude labor, and obviously the amount of pig iron which a workman might load in a day depended upon his physical strength and how he used his strength. If he worked rapidly, with little rest, he soon became exhausted. If he rested too frequently, he wasted his time. A workman's efficiency depended upon his rate of working and the length and distribution of his rest periods. The principles of scientific management were applied to the process, with the result that one workman loaded forty-seven and one half tons a day and the length of the working day was shortened.

These two illustrations are typical of a large number of instances in industrial activities where the efficiency of workmen has been greatly increased by the application of the

principles of scientific management. They suggest that the instruction in our schools may be made more efficient by the application of the same principles. For the field of education the principles of scientific management involve (1) an analysis or diagnosis of the teaching situation, and (2) the selection of methods and devices of instruction to meet the situation revealed.

Diagnosis of the teaching situation. In the consideration of the problem of measurement, it was shown that there are as many specific abilities involved in performing the operations of arithmetic as there are types of examples. These operations must be performed with a minimum of attention, so that the focus of the attention may be devoted to the consideration of problems. Thus, the teacher has the problem of engendering in each of her pupils a large number of automatic abilities or specific habits.

The individual differences of pupils furnish another factor of the teaching situation. Pupils differ in native ability and in their past experience. Some pupils are eye-minded, some are ear-minded, and still others are motor-minded. These differences become prominent in their learning. Some pupils grasp quickly the response which is to be made by seeing another perform it, others require a detailed explanation, and still others progress most rapidly by being allowed to reason out the appropriate response. Pupils also differ in the amount of practice they require to reach a given degree of facility in performing an operation.

These two conditions make the teaching situation which the teacher faces a complex one. Before she can intelligently direct her efforts as an instructor she must diagnose the

situation.¹ The tests described in this chapter furnish a means for doing this. The scores reveal the shortcomings of classes and of individual pupils.

Pupils' and class records charted. Courtis has devised a number of simple charts which can be used to exhibit graphically a pupil's record, compared with the standard scores. One form of these is shown in Fig. 3 which gives the record made by a sixth-grade pupil with the Courtis Standard Research Tests, Series B. The heavy line drawn across the chart shows the individual standards set by Courtis; the dotted line, the scores made by the pupil. A comparison of the actual scores made by the pupil with the heavy line reveals at once the examples on which this pupil needs instruction. In some of the fundamental operations he has been drilled until he is ahead of the standard for his grade. Fig. 13 in Chapter VIII shows another form of score card, and exhibits a pupil's entire record not only in arithmetic, but in reading and writing as well. The scores obtained by giving the Cleveland-Survey Tests are interpreted in the same way, and yield a more detailed diagnosis of the arithmetical abilities of a pupil.

In a similar way the strength or weakness of a class or a school may be shown by comparing the class or school medians with the standards set for the class or with the median results in the same test in other classes or schools. This is well shown in Fig. 4, which compares the median scores obtained for the whole city by the Cleveland survey

¹ Additional suggestions for diagnosing the shortcomings of pupils are given in the *Teacher's Manual for the Courtis Standard Practice Tests* (World Book Company, Yonkers, New York), and by Stone, C. W., *Standardized Reasoning Tests in Arithmetic and How to Utilize Them* (Teachers College Contributions to Education, no. 83, 1916).

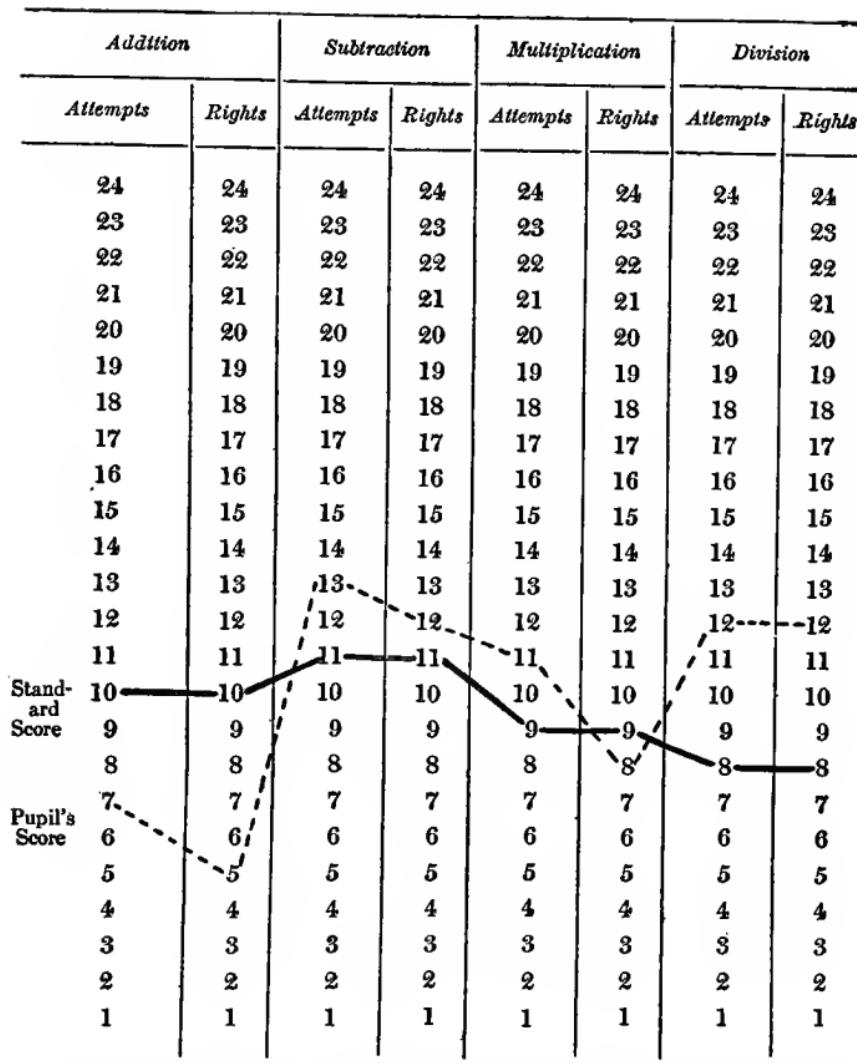


FIG. 3. A CHART, SHOWING THE SCORES MADE BY A SIXTH-GRADE PUPIL, IN COMPARISON WITH THE STANDARD SCORES, USING THE COURTIS ARITHMETIC TESTS

on the D test in simple division (see page 27) with the median scores in five selected schools. The need for closer supervision is at once evident.

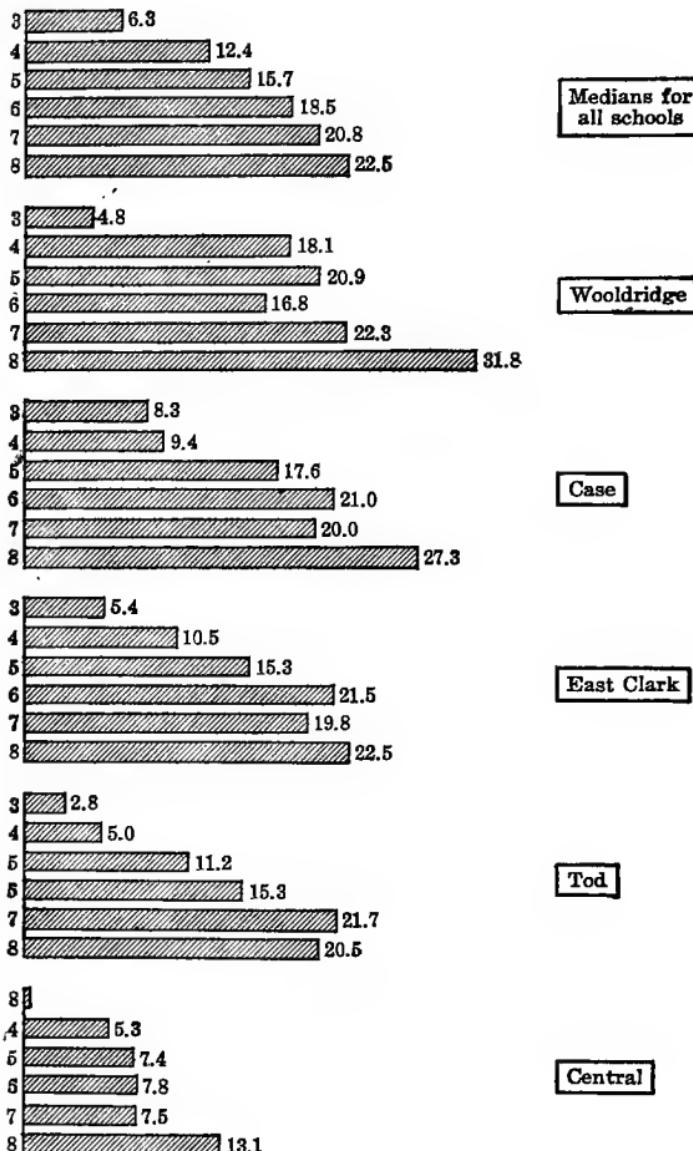


FIG. 4. MEDIAN SCORES FOR ALL CLEVELAND SCHOOLS AND FOR FIVE SELECTED SCHOOLS

(Test D, Division. See page 27.)

The Woody Arithmetic Scales also furnish a diagnosis of the class when the scores are tabulated as shown in Table II.

A more detailed study of the errors of those pupils who are below standard is needed to ascertain the causes of their weaknesses. A method¹ which has been used with success is to have the pupil being studied do the examples orally and note the errors which he makes. This gives the teacher a clearer understanding of the pupil's mental processes. By means of this method it has been found that "a large number of errors is due to the incomplete automatization of the simple facts of addition, subtraction, multiplication, and division."

Meeting the situation: Laws. In meeting the situation revealed in the case of the operations of arithmetic, the first prerequisite is that the general method of instruction be one suited to engendering specific habits or automatic responses. The laws governing the engendering of specific habits have been quite definitely established.

Stated in psychological terms, *the first law* is that in the beginning the attention of the learner shall be focalized upon the habit to be acquired. In terms of schoolroom practice this means that the learner shall understand what reaction is to be made to a given stimulus, and shall then react to it in the appropriate manner. This gives the learner the right start.

The second law is that the accomplishment of the step outlined in the first law shall be followed by attentive repetitions. It is not sufficient that there be simply repeti-

¹ Smith, James H. "Individual Variations in Arithmetic"; in *Elementary School Journal*, vol. 17, p. 195.

tions or drill. The drill must be attentive. In the case of the operations of arithmetic this drill may be detached from the solving of problems, or it may be given in the solving of problems.

The third law states that no exception shall be permitted until the habit is firmly established, which means that the attentive practice must be continued until the operation has become a habit, that is, has been made automatic.

The instruction based upon these laws must be adapted not only to the needs of the class, but also to the needs of the individual pupils which the tests reveal. The class needs can be met by placing emphasis upon the types of examples which the pupils as a group are unable to do with standard ability. This emphasis may mean simply more drill, or it may be that the difficulty is due to the pupils' not understanding how the operation is to be performed. If the latter is the case, explanation, or illustration, or opportunity to think it through is needed.

In order to be most effective, the repetitions must be attentive. This means that the drill must be effectively motivated. Arithmetic is one of the best liked of the school subjects. This is particularly true of the operations. This being the case, the motivation of drill in arithmetic is a comparatively simple matter, and in most cases it will be sufficient simply to start the pupils to work and to keep the work from lagging. When more than this is necessary the teacher must demonstrate her resourcefulness by providing an effective method or device for the motivation of arithmetical drill. In the lower grades the playing of certain games provides practice upon certain types of examples. In the

upper grades ciphering matches, or, better, the setting of definite standards in both speed and accuracy, are very effective motives.

Individual vs. class needs. However, classes are composed of individual pupils who differ in their needs. Only a few of their needs will be common to the class as a whole. The usual class instruction in arithmetic does not meet these needs. Frequently the writer has visited classes in arithmetic which were being drilled upon the fundamental operations. A fairly uniform procedure was followed. The same example was dictated to all of the pupils, regardless of whether they needed drill upon this particular type of example or not. Naturally some pupils finished very quickly, and, as they waited for their classmates to finish, there was a tendency for them to become disorderly — a perfectly natural tendency. When a majority of the class had finished the example the teacher stopped the work and read the correct answer. The process was then repeated. The result was that those pupils who worked slowly completed few, if any, examples during the entire period, and, therefore, received little satisfactory drill. The bright pupils spent a considerable proportion of their time waiting on the other members of the class, and probably did not need the particular kind of drill which they received.

The scores of a group of pupils do not cluster closely about the median. When the distributions of the scores for successive grades are compared a great overlapping is found. Some pupils in the fourth grade make higher scores than a number of the eighth-grade pupils. Table VIII shows the distribution of pupils in a certain city according to the

number of examples attempted in the subtraction test of the Courtis Standard Research Tests, Series B. An examination of this table reveals these facts.

In the fourth grade 23 per cent of the pupils reach or exceed the fifth-grade median.

In the fifth grade 23 per cent of the pupils reach or exceed the sixth-grade median.

In the sixth grade 24 per cent of the pupils reach or exceed the seventh-grade median.

In the seventh grade 40 per cent of the pupils reach or exceed the eighth-grade median.

This condition is merely typical. It is due in part to native individual differences and in part to training. In one experiment¹ three types of drill were used:—

(1) Class drill supplemented by individual assistance on the points of weakness as diagnosed by the results of the test; (2) class drill with extra drill periods provided for the slow pupils, who were drilled in groups rather than individually; and (3) merely class drill with explanations to the class as a whole.

After these types of drill had been used for a month and the results carefully measured, the following conclusions were reached:—

(1) All three types of drill produced very large increases in the achievements of the pupils. (2) Class drill supplemented by individual help at the points of weakness as diagnosed by the first test proved much more efficient on account of the exceptional decrease in the variation among the members of the class. This decrease in variation was shown by the decrease in the quartile coefficient of deviation. (3) It has been shown in both the first and second types of drill that individual variations which some writers ascribe to hereditary influences may be greatly modified by appropriate instruction.

¹ Smith, James H. "Individual Variations in Arithmetic"; in *Elementary School Journal*, vol. 17, p. 195.

TABLE VIII.

SHOWING THE DISTRIBUTION OF THE PUPILS OF A CITY ACCORDING TO THE NUMBER OF EXAMPLES ATTEMPTED, COURTS STANDARD RESEARCH TESTS, SERIES B

Grade	Number of examples attempted										Median
	1	2	3	4	5	6	7	8	9	10	
IV	1	7	11	17	19	18	10	8	4	4	1
V		1	3	9	10	18	23	21	8	2	3
VI			1	2	8	9	20	16	14	9	1
VII				1	2	6	9	13	12	16	14
VIII					1	1	9	13	8	19	12

Obviously, it is the pupil who performs the operation slowly and with difficulty who needs practice. The pupil who already is skillful in performing the operation does not need further drill. Our present procedure provides drill for the pupil who does not need it, and prevents the pupil who does need it from receiving it in a satisfactory manner.

Repeating the tests after an interval. The effect of class instruction is shown when a test is repeated after an interval of a few months. In Table IX the midyear¹ distribu-

TABLE IX. THE RANGE OF NUMBER OF EXAMPLES ATTEMPTED

Grade	Addition			
	Total Range	Range in number of examples	Range of middle 50 per cent	Range in number of examples
IV.....	1-10	10	3.6- 5.9	2.3
	1-17	17	5.0- 8.3	3.3
V.....	2-14	13	4.7- 7.0	2.3
	3-16	14	6.8- 9.0	2.2
VI.....	2-12	11	5.7- 8.4	2.7
	4-24	21	7.2-11.0	3.8
VII.....	3-24	22	6.9-10.0	3.1
	4-24	21	8.5-12.3	3.8
VIII.....	4-18	15	8.1-10.8	2.7
	5-17	13	8.9-12.1	3.2

For each grade the upper line is for January and the lower line is for May.

tions for a certain city on the addition test of Series B are compared with those for May. With only two exceptions both the total range of the scores and the range of the middle fifty per cent were increased. This fact shows that the in-

¹ The first test was given just after the midyear promotions. The tests were given to about one hundred and fifty pupils in each grade.

struction in addition which these pupils received was more appropriate for the brighter pupils than for those who were below standard ability. Some pupils acquired abilities far in excess of the standard for their grade, while others remained conspicuously below the standard. This is merely what we should expect, because those pupils who have profited most under our system of instruction may be expected to continue to profit most. Obviously, if our standards are wisely determined, the pupils who are below standard in ability need instruction and those who are conspicuously above standard may spend their time more wisely upon other subject-matter. If this is not feasible, the methods and devices of instruction should be those most appropriate to those pupils who are below standard. If the methods of instruction are unchanged, it is obvious that the pupils who have learned most readily will continue to do so.

The bright pupil should receive consideration as well as the backward pupil. The usual class instruction does not give the bright pupil efficient training. He is not forced to exert himself. Much of the time he is forced to be inactive. Furthermore, in the case of the tool subjects (the operations of arithmetic, reading, spelling, handwriting and language) training beyond a certain point is not very profitable. In arithmetic the bright pupil should be given problem work rather than additional training upon the operations.

Modifying the class drill. The type of class instruction described on page 55 can easily be modified so as to insure that the slow-working pupils will get some satisfactory drill. Instead of dictating only one example at a time, the teacher can dictate several, and stop the work as soon as a few of the

faster workers have finished. The slow-working pupils will have some examples completed.

The teacher must recognize that the rate at which the pupil performs the operations is important, as well as the accuracy. This means that in teaching the teacher must obtain a measure of the pupil's speed, as well as a measure of his accuracy. If examples are dictated in groups, and the work stopped as suggested in the above paragraph, the number of examples which the pupil does during the class period is a measure of his rate of working. The per cent correct is a measure of his accuracy.

The instruction can be made still more effective if the teacher will prepare a number of sets of examples, each set being confined to examples of the same type. These sets of examples should be written on cards. Then, instead of dictating examples, the teacher can distribute the cards and have the pupils copy the examples from the cards. If the teacher studies the needs of her pupils, it will be possible for her to distribute the cards so that each pupil will have the type of example upon which he needs practice. The pupil is probably injured by being required to practice upon the wrong type of example and, hence, it is very important that each pupil be given the type of example upon which he needs practice.

Use of practice tests. Courtis has devised a set of Standard Practice Tests¹ which automatically diagnoses each pupil and furnishes the practice which he needs to remedy his defects. These tests consist of forty-eight sets of exer-

¹ Full details regarding these tests may be obtained from the publishers, World Book Company, Yonkers, New York, and Chicago, Illinois.

cises, which "have been designed to cover every known difficulty in the development of ability in the four operations with whole numbers." The latest form of these tests (1916) is arranged so that the pupils begin the series by taking Lesson 13, a test involving all types of examples found in the first twelve lessons.¹ All pupils who attain standard ability on this test are excused from the first twelve lessons, because they have demonstrated that they do not need the instruction which these lessons provide. As soon as a pupil who did not attain standard ability on Lesson 13 has finished the first twelve lessons, he takes Lesson 13 again to show that he is now up to standard. Lessons 30, 31, and 44 are also test lessons, and are used in the same way.

Each of the lessons is printed upon a card and a copy is furnished to each pupil. The card is placed beneath a sheet of transparent paper and the example is read through the paper, the work being done on the paper. The lessons have been constructed so that the standard length of time required to complete each one is the same. They are also self-scoring. These two features relieve the teacher of the laborious work of scoring the papers, and make it possible for different pupils to be working upon different lessons at the same time. Thus, when a pupil has demonstrated that he is up to standard on any type of example, he may at once go on to the next lesson. If he is not up to standard on any lesson his work makes the fact obvious, and he can remain upon that lesson until he acquires the necessary

¹ All lessons except the test lessons are confined to a single type of example.

ability without interfering in the least with the work of the other members of the class.

Thus, individual progress is provided for, and at the same time the group formation is retained. A considerable saving of pupil's time is effected by excusing from drill those pupils who demonstrate that they possess standard ability. These pupils can spend this time upon other work.

These "Standard Practice Tests" also simplify instruction in ungraded schools. The same lessons are used for all pupils in grades four to eight. Only the time allowed differs. Thus all of the pupils in a rural school could be instructed at the same time and each pupil receive the practice which he needed.

Another series of exercises, known as the "Studebaker Economy Practice Exercises," and based upon some of the same general principles, has been devised by J. W. Studebaker, Assistant Superintendent of Schools, Des Moines, Iowa. They are published by Scott, Foresman & Company, New York and Chicago. Other series of practice exercises have been devised, but, so far as the writer has examined them, they are less complete and give less promise of efficient means of instruction.

However, it must not be forgotten that any set of practice exercises are merely teaching devices. It is more important that the teacher explicitly recognize in her thinking that she is instructing a group of pupils who differ widely in native ability, experience, and training, that all do not learn in the same way and that a limitation should be placed upon training. When she explicitly recognizes these facts, the resourceful teacher will find many devices which will be

helpful in adapting the instruction to the needs of the pupils.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. Which of the first three tests described in this chapter would you select in order to secure the most helpful diagnosis of the class and of the individual pupils? Why?
2. How can the tests described in this chapter be used by the teacher to make her instruction more effective?
3. Do you think pupils will welcome definite objective standards and the use of standardized tests? Why?
4. If you are using standardized tests make charts showing class (or individual) scores in comparison with the standards. Some teachers have found it helpful to have such charts hung in the classroom. It is also helpful to bring such charts to the attention of the patrons of the school.
5. Make a chart showing how the pupils of your class compare with other classes of the same grade and with classes of other grades.
6. Suppose a pupil is unable to do satisfactorily certain types of examples. How would you proceed to locate his particular difficulties? If you are teaching arithmetic try out your plan on some of your pupils.
7. What devices do you use to provide each pupil with the training which he needs? What devices are suggested in this chapter? Can you suggest additional ones?
8. Pupils who are excused from drill because they do not need it should spend their time doing profitable things. Suggest a number of assignments which might be made to such pupils. The assignments may be in subjects other than arithmetic if it seems wise, but they should be such as not to interfere with the instruction of the other pupils.
9. How do you know that the methods and devices of instruction which you are now using are the best? How could you find out?
10. How do you know that you are not giving too much time to arithmetic? How could you find out?
11. Is a class score which is conspicuously above standard a sign of superior teaching? Why?
12. Construct two tests, each being confined to a single type of example. Give both tests to the same pupils under the same conditions. Compare the two sets of scores.
13. Scientific experimentation will be necessary to determine the best plans of grouping pupils for instruction. These plans are worthy of a trial.
 - a. In a building place together for drill those pupils which are most nearly equal in ability as shown by the tests.

- b. Excuse from drill those who have demonstrated that they are above standard.
- c. Have a special "hospital" class for those pupils who have scores materially below standard. A pupil's sentence to the "hospital" would be until he was up to standard.

BIBLIOGRAPHY

Only the most important references are given here. Additional references are given in footnotes in the chapter.

I. TESTS IN THE FUNDAMENTAL OPERATIONS

1. *Courtis Standard Research Tests, Series B.* The tests may be secured from S. A. Courtis, 82 Eliot Street, Detroit, Michigan.

REFERENCES: Courtis, S. A. *Third, Fourth, and Fifth Annual Accountings, 1913-1916.* (Department of Coöperative Research. 82 Eliot Street, Detroit, Michigan.)

Ashbaugh, E. J. *The Arithmetical Skill of Iowa School Children.* (Bulletin no. 24. Extension Division, University of Iowa.)

Ballou, Frank W. *Educational Standards and Educational Measurements, with Particular Reference to Standards in the Four Fundamentals in Arithmetic.* (School Document no. 10. 1914. Boston Public Schools. Bulletin no. 3, Department of Educational Investigation and Measurement.)

Haggerty, M. E. "Arithmetic: A Coöperative Study in Educational Measurements." (Indiana University Studies, no. 27.)

Haggerty, M. E. (editor). *Studies in Arithmetic.* (Indiana University Studies, no. 32, vol. 3. September, 1916.)

Monroe, Walter S. *A Report of the Use of the Courtis Standard Research Tests in Arithmetic in Twenty-four Cities.* (Kansas State Normal School, Emporia; Bulletin, new series, vol. 4, no. 8.)

2. *Research Tests in Arithmetic, Addition of Fractions, designed by F. W. Ballou.* Copies of these tests are not obtainable.

REFERENCE: *Arithmetic.* (Bulletin no. 7, Department of Educational Investigation and Measurement, Boston.)

3. *The Cleveland Survey Arithmetic Tests.* Copies of the test papers may be obtained from Charles H. Judd, School of Education, University of Chicago, Chicago, Illinois.

REFERENCES: Judd, Charles H. *Measuring the Work of the Public Schools.* (Cleveland Foundation, Survey Report, Cleveland, Ohio.) Also for sale by the Russell Sage Foundation, New York City.

Smith, James H. "Individual Variations in Arithmetic"; in *Elementary School Journal*, vol. 17, pp. 195-200.

4. *Stone's Arithmetic Test for the Fundamental Operations.* Designed as a general test. Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCE: Stone, C. W. *Arithmetical Abilities and Some Factors Determining Them.* (Teachers College Contributions to Education, no. 19.) Obtained from above address.

5. *Arithmetic Scales devised by Clifford Woody.* Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCE: Woody, Clifford, *Measurements of Some Achievements in Arithmetic.* (Teachers College Contributions to Education, no. 80.) Obtained from above address.

II. REASONING TESTS

1. *Stone's Reasoning Test.* For copies of the test, address Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCES: Stone, C. W. *Arithmetical Abilities and Some Factors Determining Them.* (Teachers College Contributions to Education, no. 19.) Obtained from above address.

Stone, C. W. *Standardized Reasoning Tests in Arithmetic and How to Utilize Them.* (Teachers College Contributions to Education, no. 83.)

2. *Starch's Arithmetical Scale A.* Copies may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.

REFERENCE: "A Scale for Measuring Ability in Arithmetic"; in *Journal of Educational Psychology*, April, 1916.

3. *Buckingham's Reasoning Test in Arithmetic.* Used by Buckingham in the Survey of the Gary, and the Prevocational Schools of New York City.

REFERENCES: *Seventeenth Annual Report of the City Superintendent of Schools*, New York City, 1914-15.

Buckingham, B. R. "Notes on the Derivation of Scales, with special reference to Arithmetic"; in the *Fifteenth Yearbook of the National Society for the Study of Education*, part I.

CHAPTER III

READING

The problem of measurement in reading. The problem of measurement in reading differs in certain important respects from that in arithmetic. An arithmetical example calls for a definite response from the pupil. All other responses are incorrect. Furthermore, our arithmetical symbols and number system are such that it is comparatively easy for the pupil to give unmistakeable objective evidence of the character of his response. In the case of reading, particularly silent reading, the appropriate response to a sentence or paragraph is not so clearly defined, and it is not easy for the pupil to give objective evidence of the response which he has made. The several attempts to solve the problem of measurement for reading are described in the following pages.

I. SILENT READING

Silent reading may be classified into two main divisions, for each of which tests have been prepared. The *first* deals with the ability of the pupil to read and know the meaning of the words he reads (*a*) and the *second* deals with the rate of reading and the degree of comprehension shown (*b*). The first is essentially vocabulary; the second is essentially understanding and speed.

(a) **Recognition of words.** A fundamental factor of one's ability to read silently is the range of words whose meaning

he recognizes. Three tests have been devised to measure a pupil's ability to associate the appropriate meanings with printed words.

1. The Thorndike Visual Vocabulary Scales¹

Thorndike is the author of three visual vocabulary scales: Scale A, Scale A₂, and Scale B. The latter two represent extensions of the former, and were derived by the same method. Scale A₂ and Scale B are intended for use alternately or interchangeably, and so a brief description of Scale A₂ only will be given here.

The scale consists of twenty-three lists of words. Each list contains ten words. The words in any given list possess about equal difficulty as to their meaning. For ability to indicate the meanings of the words comprising any given list a certain score is given, and for ability to indicate the meanings of the words in the more difficult lists a correspondingly higher score is assigned. The score values of the lists of words were determined by having several thousand children undertake to indicate the meaning of each word. The greater the per cent of children who could not give the meaning of the word, the higher the score value attached to the word. The method used in giving the tests can be made clear by quoting from the opening lines of the test sheet:—

Write the letter F under every word that means a flower.

Write the letter A under every word that means an animal.

Write the letter N under every word that means a boy's name.

Write the letter G under every word that means a game.

¹ Thorndike, E. L. "The Measurement of Ability to Read"; in *Teachers College Record*, September, 1914, for Scale A. *Teachers College Record*, November, 1916, for Scale A₂ and Scale B.

And so on for four additional meanings. Then the twenty-three lists of words follow, with the score value of each list given on the left-hand margin.

Three representative lists are reproduced below:—

Value 4. daisy camel samuel rabbit monkey william
tulip goat paul violet

Value 7. constantly sincere chess weak antelope eugene
henceforth julian formerly candy-tuft

Value 10. phlox set dependable judicious caribou orchid
ruthless compassionate cyrus petunia

Use and standards. The author of the scale recommends that, in using it, the examiner should not rate the child at the highest point in the scale, where he knows all the words in the list, but rather at that point where he knows only eight out of ten. A table of values is provided by which to infer the proper score where the child misses either more or less than two words in the list which best measures his acquaintance with words.

Complete instructions for giving and scoring the test, preliminary sheets, and scoring sheets are provided with the tests. Permission is also granted to any one wishing to use the tests to print them for himself, in which case it is recommended that brief test sheets be made up from the scales restricting the lists used with each grade of children to a few values. Thus much time will be saved, both for children and teacher, by eliminating the difficult lists for the younger children and the easy lists for the older children.

No standards have as yet been derived by the use of the Thorndike Scale A₂ or Scale B with large numbers of public school children. In Table X the standards of achievement

by the use of the Thorndike Scale A (with which the values on scales A₂ and B are supposed to be identical) are given, and serve as tentative standards for purposes of comparison. The score values were obtained by the measurement of the pupils in eighteen cities in Indiana.¹

TABLE X. MEDIAN SCORES IN VISUAL VOCABULARY BY THE THORNDIKE SCALE A

	<i>Grades</i>					
	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Median score.....	4.00	5.26	6.00	6.66	7.29	7.91
Number of children.....	1650	2095	2028	1860	1625	1313

2. *The Haggerty Visual Vocabulary Tests*

The tests prepared by Haggerty, of the Bureau of Coöperative Research of the School of Education of the University of Minnesota, are but a slight modification of the Thorndike scales described above, with the addition of an oral test for children of Grades I and II. This test will be described under the heading of "Oral Reading."

Scale R₂, of which there is one sheet for children of Grades III and IV, and another sheet containing part of the same words and additional more difficult words for Grades V, VI, VII, and VIII, are devised in exactly the same way as the Thorndike scales. Methods of scoring are somewhat more simple, and the lists are more brief than those used by Thorndike. Standards are being worked out in the Minnesota schools for each grade, but as yet none have been announced.

¹ Haggerty, M. E. *The Ability to Read: Its Measurement and Some Facts Conditioning It.* (Indiana University Studies, no. 34.)

3. Starch's English Vocabulary Tests¹

These tests are lists of one hundred words, each selected at random from a dictionary. The child is asked to check the words of the meaning of which he is certain, and to write the meaning after the words where he is in doubt. The child's score is the per cent of words thus checked or correctly defined.

To illustrate the Starch vocabulary lists, the first one of his two is reproduced below.

STARCH'S ENGLISH VOCABULARY TEST-LIST I

1. acta	24. currency	47. to interlay
2. abnormal	25. death	48. Italianate
3. agriculturist	26. departmental	49. Jupiter
4. ambulacrum	27. difference	50. knowledgeable
5. Araneida	28. displayed	51. Latin
6. assagia	29. to dow	52. lewis
7. awaft	30. dysodile	53. loam
8. barker	31. eloquence	54. Lycoperdon
9. belleric	32. epicine	55. mange
10. bizarre	33. evaporative	56. mayonnaise
11. bonmot	34. faction	57. mesotasis
12. drible	35. to flat	58. miscue
13. butter-cup	36. forest	59. moon
14. canon	37. fubby	60. musk
15. Catananche	38. to gazette	61. neovolcanic
16. chancroid	39. glonion	62. to notate
17. to chop	40. gyral	63. off shore
18. clearness	41. hautboy	64. organdie
19. collar	42. heterogony	65. owlet
20. to comprobate	43. hordeaceous	66. parallel
21. constructiveness	44. hyperkeratosis	67. to peal
22. to cree	45. to implore	68. personable
23. correal	46. to infatuate	69. to piece

¹ Starch, Daniel, *Educational Measurements*, p. 38.

70. Pluerotoma	81. secessionist	92. tipburn
71. portrait	82. sex	93. to transfer
72. prevailing	83. sigmoid	94. to trump
73. proveditor	84. to sluice	95. unbeseem
74. quadruple	85. spadloon	96. upholster
75. rapt	86. spur	97. vernier
76. reformer	87. stipulator	98. waldgrave
77. respectful	88. subregion	99. wharf
78. river	89. sweet	100. zelotypia
79. rutter	90. tarsus	
80. sawmill	91. Theatin	

(b) **Tests for comprehension and speed.** There are three types of tests to measure the ability of pupils to read silently sentences or series of connected sentences. *First*, tests for the understanding of sentences or paragraphs, without regard to the time required for that understanding. In this class are the Thorndike Scale Alpha, and the Minnesota Scale Beta. *Second*, tests which measure separately the speed of reading and the amount of comprehension. In this class are the Courtis English Tests, Brown's Silent Reading Test, Starch's Silent Reading Tests, and Gray's Silent Reading Tests. *Third*, tests which combine the factors of speed and comprehension in a single mark or score. In this class are the Kansas Silent Reading Tests and the Courtis Silent Reading Tests.

1. *The Thorndike Scale Alpha*¹

This scale is to measure the understanding of sentences. It consists of a series of groups of sentences which the child is to read. These groups of sentences or paragraphs represent increasing degrees of difficulty, the extent of which has

¹ Thorndike, E. L. "An Improved Scale for Measuring Ability in Reading"; in *Teachers College Record*, November, 1915, and January, 1916.

been determined experimentally. This difficulty is expressed by an appropriate number opposite each paragraph. Each paragraph is followed by a number of questions which test the child's understanding of what he has read. The element of speed of reading is disregarded, the child being allowed all the time he requires. The following set of sentences and questions upon them are copied from the scale, and illustrate the sort of ability called for.

Set C (Score Value, 8)

Read this and then write the answers. Read it again as often as you need to.

It may seem at first thought that every boy and girl who goes to school ought to do all the work that the teacher wishes done. But sometimes other duties prevent even the best boy or girl from doing so. If a boy's or girl's father died and he had to work afternoons and evenings to earn money to help his mother, such might be the case. A good girl might let her lessons go undone in order to help her mother by taking care of the baby.

1. What are some conditions that might make even the best boy leave school work unfinished?.....
2. What might a boy do in the evenings to help his family?.....
3. How could a girl be of use to her mother?.....
4. Look at these words: *idle, tribe, inch, it, ice, ivy, tide, true, tip, top, tit, tat, toe.*

Cross out every one of them that has an *i* and has not any *t* (T) in it.

A key for scoring is provided by which the examiner is guided in deciding which answers to score right and which to score wrong. The child's score is the score value of the most difficult paragraph concerning which he can answer eighty per cent or more of the questions. The class score is

the score value of the most difficult paragraph concerning which the per cent of correct answers made by all the class is nearest eighty. A table of values is provided for interpolating where the per cent is somewhat above or below eighty. Table XI gives the median scores for the pupils in eighteen cities in Indiana as reported by Haggerty.

TABLE XI. MEDIAN SCORES IN UNDERSTANDING OF SENTENCES BY THE THORNDIKE SCALE ALPHA

	<i>Grades</i>					
	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Median.....	5.48	6.56	7.56	8.46	8.72	9.00
Number of pupils.....	1650	2095	2028	1860	1625	1313

2. The Minnesota Scale Beta

The Bureau of Coöperative Research of the School of Education of the University of Minnesota has had printed a slightly modified form of the Thorndike Scale Alpha under the name Scale Beta. One sheet is prepared for Grades III to V, and another sheet for Grades VI to IX. Certain exercises are common to both sheets, but the most difficult ones are omitted for the younger children, and the least difficult ones are omitted for the older children. The method of scoring is the same as used with the Thorndike Scale Alpha.

3. The Courtis English Tests¹

As an example of one of the first attempts made to measure speed of reading separately from comprehension, mention

¹ Courtis, S. A. *Manual of Standard Tests*; also, *The Fourteenth Yearbook of the National Society for the Study of Education*, part I, pp. 44 to 56.

should be made of the Courtis English Tests, a part of which involves measuring the number of words read per minute when the child knows that he is going to be asked to reproduce the meaning, and also the number of words read per minute when no reproduction is to be called for. The test also provides for a very careful measure of comprehension as determined by the amount of the meaning which the child can reproduce in composition form. The task involved in giving the tests and scoring the papers is so laborious, however, as practically to nullify the virtues of the tests because none but the most enthusiastic investigators are willing to undertake to use them. Their chief service has been as a type after which other much more simplified tests have been modeled.

4. Brown's Silent Reading Test¹

This test consists of a very interesting reading selection. The directions require that the children being tested read the selection silently exactly one minute, then draw a line around the word which they have reached when the examiner calls "stop." The number of words read makes the score in speed.

The children are then asked to write as much as they can remember of what they have read. A key is provided for the examiner to use in scoring the papers. On the key is listed all the separate ideas contained in the selection. By comparing the child's papers with the key, the examiner

¹ Brown, H. A. *The Measurement of Ability to Read*. (Bulletin no. 1, Bureau of Research, State Department of Public Instruction, Concord, New Hampshire.)

determines first how many different points there are in what the child read. Then his reproduction is examined carefully to determine (1) quantity and (2) quality of comprehension. Quoting from the *Manual*, page 15:—

A good method of procedure in scoring the papers is to place the key and the child's reproduction side by side on a table. The first point in the key should be examined and then the child's reproduction should be carefully scrutinized to see if it contains the idea. It should be borne in mind that the child's language may be entirely different from the wording of the idea in the key. The purpose of the one scoring the paper should be to see if that idea is expressed in the child's own language or sufficiently plainly implied by what he has written with sufficient completeness and accuracy to be credited for quantity. If so it should be scored for quantity. Each point should be considered in the same manner. Each idea should next be examined with reference to quality and scored if it meets the requirements. When the entire paper has been scored, the number of ideas scored for quantity and the number for quality should be counted and each expressed as a percentage of the total number in the portion of the selection read by the particular child in question. The average of quantity and quality should be taken as a measure of the child's comprehension.

The scorer must use his best judgment in determining whether to give credit for an idea. Everything which the child says in any way related to the idea in question should be considered, and if it seems from his language that he has the idea with the required degree of completeness and correctness in the case of either quantity or quality, it should be credited.

The record sheet calls for the reduction to one mark, termed "reading efficiency," of the two marks of speed and comprehension by multiplying them together. Thus, if a child reads 1.5 words per second, and the average of his quantity and quality of comprehension is 64, his reading efficiency is 1.5 times 64, or 96.

Value of the test in diagnosis. For purposes of diagnosis

this test has marked advantages. Not only can the teacher determine the standing of his class in general reading efficiency, but the arrangement of the data on his class record sheet enables him also to see at a glance whether his class or any individual child in it is up to standard in speed, in quantity, or in quality of comprehension. Several equivalent selections have been prepared so that it is possible to use new material when the test is repeated.

Brown proposes as tentative standards, for purposes of comparison, the average scores made by the best class yet tested in each school grade. These averages are given in Table XII.

TABLE XII. TENTATIVE SCORES WITH THE BROWN SILENT READING TEST

	<i>Rate of reading</i>	<i>Comprehension</i>	<i>Reading efficiency</i>
Grade III	3.32 words per second	46	127.8
Grade IV	3.55 " " "	65	217.1
Grade V	4.40 " " "	61	291.0
Grade VI	4.54 " " "	68	295.0
Grade VII	4.65 " " "	87	322.3
Grade VIII	4.84 " " "	79	323.6

5. Starch's Silent Reading Tests¹

These tests do not differ essentially from Brown's test, described above. Instead of using one selection for all ages of children, the Starch tests consist of a different selection

¹ Starch, Daniel. "The Measurement of Efficiency in Reading"; in *Journal of Educational Psychology*, January, 1915. Also in his *Educational Measurements*, pp. 22-30.

for each grade in the elementary school. The following one, which is used for fifth grade, will illustrate:—

No. 5

Once upon a time, there lived a very rich man, and a king besides, whose name was Midas; and he had a little daughter, whom nobody but myself ever heard of, and whose name I either never knew, or have entirely forgotten. So, because I love odd names for little girls, I chose to call her Marygold.

This King Midas was fonder of gold than anything else in the world. He valued his royal crown chiefly because it was composed of that precious metal. If he loved anything better, or half so well, it was the one little maiden who played so merrily around her father's footstool. But the more Midas loved his daughter, the more did he desire and seek for wealth. He thought, foolish man! that the best thing he could possibly do for his dear child would be to give her the immensest pile of yellow, glistening coin, that had ever been heaped together since the world was made. Thus, he gave all his thoughts and all his time to this one purpose. If ever he happened to gaze for an instant at the gold-tinted clouds of sunset, he wished that they were real gold, and that they could be squeezed safely into his strong box. When little Marygold ran to meet him with a bunch of buttercups and dandelions, he used to say, "Poh, poh, child! If these flowers were as golden as they look, they would be worth the plucking!"

And yet, in his earlier days, before he was so entirely possessed of this insane desire for riches, King Midas had shown a great taste for flowers.

The children being tested are allowed just thirty seconds to read as much as they can. When the time is up they make a mark after the last word read, and then turn the sheet over and reproduce as much as possible of what they have read. It is recommended that the test be repeated the succeeding day, using the selection designed for the grade below the one the child is in.

The average number of words read per second in the two

tests makes the child's score in speed. For the score in comprehension, the child's reproduction is carefully examined, and all words crossed out which do not represent ideas contained in the test passage, or which repeat ideas recorded. The remaining words are counted and the average of the numbers of words in the two reproductions is used as the score in comprehension.

The author of the tests recommends that when large classes of children are being tested, and the results are not intended for use with individuals, the reproductions be not examined but that the whole number of words used be counted and reduced by seven per cent for errors. When thus treated the tests become wholly objective.

The median scores made by six thousand children in twenty-seven schools are given in Table XIII.

TABLE XIII. MEDIAN SCORES IN READING BY STARCH TESTS,
ATTAINED AT THE CLOSE OF THE RESPECTIVE YEARS

<i>Grades or years</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>
Speed of reading (words per second) .	1.5	1.8	2.1	2.4	2.8	3.2	3.6	4.0
Comprehension (words written)	15	20	24	28	33	38	45	50

6. Gray's Silent Reading Tests¹

These tests consist of three selections, one for Grades II and III, one for Grades IV, V, and VI, and another for

¹ Gray, W. S. "Tests of Silent Reading"; in Charles H. Judd, *Measuring the Work of the Public Schools*, pp. 275-81. (The Cleveland Foundation Survey, 1916.) For other references see Bibliography at end of chapter.

Grades VII and VIII. The selections are so arranged on the pages that the time required to read one hundred words can be readily ascertained. Only one child is tested at a time. After completing the reading, the child, if in the second or third grade, tells the story to the examiner who writes it down. If in the grades above the third, the child writes all he can remember of the story, and then writes answers to a set of questions which is furnished him by the examiner. The child's score for quality of reading is assigned on the basis of two factors, reproduction and accuracy. Reproduction is determined by the number of words which remain in the child's composition, after all wrong or irrelevant statements and repetitions are stricken out. Accuracy is determined on the basis of ten points for each correct answer. The quality mark is the average of these two. Thus a score for rate of reading and a score for quality of reading is ascertained for each pupil. Standards are given in Table XIV.

TABLE XIV. STANDARD SCORES FOR GRAY'S SILENT READING TESTS

<i>Grade</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>
Rate (words per second).	1.50 32	2.30 37	2.20 29	2.57 32	2.79 39	2.69 22	2.87 27
Quality.....							

7. *The Kansas Silent Reading Tests*¹

The Kansas Silent Reading Tests take both speed and accuracy of comprehension into account in a single mark.

¹ Kelly, F. J. "The Kansas Silent Reading Tests"; in *Journal of Edu-*

Each test — No. 1, for Grades III, IV, and V; No. 2, for Grades VI, VII, and VIII; and No. 3 for the high school — consists of sixteen exercises, of which the following, intended for Grades VI, VII and VIII, will serve as illustrations: —

No. 1. Value, 1.0	The air near the ceiling of a room is warm, while that on the floor is cold. Two boys are in the room, James on the floor and Harry on a box eight feet high. Which boy has the warmer place? . . .
No. 6. Value, 2.3	In going to school, James has to pass John's house, but does not pass Frank's. If Harry goes to school with James, whose house will Harry pass, John's or Frank's? . . .
No. 14. Value, 4.9	A list of words is given below. One of them is needed to complete the thought in the following sentence: The roads became muddy when the snow. . . . Do not put the missing word in the blank space left in the sentence, but put a cross below the word in the list which is next above the word needed in the sentence. water is melted snow.

The supposition is that the child's comprehension is tested by each exercise. And since he does not pass on to the next exercise until he has indicated his comprehension of each one, the sum of the values attached to the exercises which he can complete correctly in the five minutes allowed is thought to give a fair measure of his speed of reading combined with his ability to comprehend the meaning of the exercises.

Educational Psychology, February, 1916; also, Bulletin no 3, Bureau of Educational Measurements and Standards, State Normal School, Emporia, Kansas.

Score values for the Kansas tests. The score to be given for the correct doing of each exercise is indicated in the margin to the left of the exercise. This value was determined by the length of time children of a given grade did actually require on the average to do each exercise correctly.

Standard Median Scores. These tests have been used widely and very reliable standards have been established for the several grades. In Table XV the median score for each grade is given, and also the twenty-five percentile, or that mark below which the poorest one fourth of the children of the grade fall, and the seventy-five percentile, or that mark above which the best one fourth of the children of the grade fall.

TABLE XV. MEDIAN SCORES, KANSAS SILENT READING TESTS
(Based upon more than 100,000 scores)

	Grade III	Grade IV	Grade V	Grade VI	Grade VII	Grade VIII	Grade IX	Grade X	Grade XI	Grade XII
Twenty-five percentile.....	2.5	6.1	9.4	9.4	11.8	13.7	16.0	17.9	18.7	22.3
Median score...	5.3	9.5	13.2	13.9	16.2	19.2	22.9	25.8	28.5	29.7
Seventy-five percentile.....	8.2	13.6	17.5	19.8	21.9	26.4	30.4	31.9	33.1	34.1

8. The Courtis Silent Reading Tests¹

These tests are of the same general type as the Kansas Silent Reading Tests. In them, however, the several exercises make a connected story. The response called for in all the exercises is the same, namely, to draw a line around

¹ Courtis, S. A. *Standard Research Tests in Silent Reading*, 82 Eliot Street, Detroit, Michigan.

a word. Furthermore, the exercises are selected with a view to their equality in respect to difficulty, and so no attempt has been made to assign values to them. The score of the child taking the test is the number of the exercises which he can complete correctly in the time allowed.

No standards have yet been obtained by the use of these tests.

II. ORAL READING

Silent reading represents one part, and quite a large part of a child's ability to use reading as a tool. Oral reading represents another part, and for this also a number of tests have been devised.

1. *The Jones Visual Vocabulary Tests*¹

Jones sought a means of measuring the vocabulary of primary grade children. Selecting ten of the most widely used primers, he found the frequency of occurrence in all the primers of each word occurring in any of them. He used this frequency as a measure of the value of each word. Thus a word occurring one hundred times would count twice as much in a child's score as a word occurring only fifty times. Using the values thus determined for each word, lists of words were made up as tests, and a child's score in the test is the sum of the values attached to the words which he can pronounce correctly.

¹ Jones, R. G. "Standard Vocabulary for Primary Grades"; in *The Fourteenth Yearbook of the National Society for the Study of Education*, part I.

2. The Haggerty Visual Vocabulary Tests¹

For Grades I and II these consist of two sheets, one of sight words and the other of phonetic words selected from the Jones test. The words on either sheet are grouped into lists according to difficulty. This difficulty was determined by trial with several hundred primary children. Values are attached to each word according to its ascertained difficulty. The child being tested is asked to pronounce the words aloud, and his score is the value attached to the most difficult list of which he can pronounce four out of five words correctly.

3. Gray's Oral Reading Test²

Nature of the tests. This test consists of eleven paragraphs, arranged in order of increasing difficulty. The relative difficulties have been established experimentally. Below are reproduced the first, sixth, and eleventh paragraphs, as illustrations.

1. A boy had a dog.
The dog ran into the woods.
The boy ran after the dog.
He wanted the dog to go home.
But the dog would not go home.
The little boy said,
 “I cannot go home without my dog.”
Then the boy began to cry.

¹ Haggerty, M. E. “Scales for Reading Vocabulary of Primary Children”; in *The Elementary School Journal*, vol. 17, no. 2. (October, 1916.)

² Gray, W. S. “Oral Reading Test”; in Charles H. Judd, *Measuring the Work of the Public Schools*, pp. 263-75. (Cleveland Foundation Survey, 1916.) See Bibliography at end of chapter for additional references.

6. It was one of those wonderful evenings such as are found only in this magnificent region. The sun had sunk behind the mountains, but it was still light. The pretty twilight glow embraced a third of the sky, and against its brilliancy stood the dull white masses of the mountains in evident contrast.
11. The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

How the test is scored. The plan of administering the test is rather complicated. The child's oral reading of each paragraph is checked for time, and for each of six types of errors; namely, gross errors, minor errors, omissions, substitutions, insertions, and repetitions. These errors are defined in the manual. The credit given for reading any paragraph varies inversely with the time and inversely with the number of errors. For example, certain credit is given a second-grade child for reading paragraph 1 in forty seconds with less than five errors, and additional credit is given the same child for reading the same paragraph in thirty seconds with less than five errors, or in forty seconds with less than four errors. Still different credit is given to third-grade children for each of the above achievements with paragraph 1. When the combination of length of time and number of errors exceeds a certain prescribed maximum, no credit is allowed. The score of any child is ascertained by adding together all the credits which he has earned on the several paragraphs. This process becomes much more simple than it sounds here when the blanks for recording

the detailed data for each child and for tabulating results are at hand.

III. AN ESTIMATE OF THE VALUE OF THE SEVERAL READING TESTS

Before passing to a discussion of the chief uses of the reading tests described above, it may be well to point out some of their distinguishing virtues and at the same time some of their limitations. Let it be remembered at the outset that all standard test work in the field of education is still in the pioneer stage, and that most authors of tests, if not all, realize more keenly than their critics how far short of perfect instruments these tests now are.

Criteria for estimating values. As a basis for judging the usefulness of the present available reading tests, four of the commonly accepted criteria for determining satisfactory educational tests in any subject are here given.

(1) The test must be objective. That is to say, there must be room for a minimum of opinion when rating the results of the test. A perfect test in this regard would be one which when answered by any child would be rated alike by all competent judges. Perfect objectivity probably cannot be hoped for in reading tests, but it is one of the essential things to be sought in a test and it is worth making considerable of a sacrifice to approach. Without objectivity it is quite impossible to use a test to compare the standards of achievement from school to school unless the papers from all the schools are rated by the same judges. Since this is not practicable, and since one of the chief uses of standardized tests is to make possible comparisons among schools, objectivity

comes to be of the first importance as an essential of a satisfactory test.

(2) The test must be arranged in steps along a scale whose units are equal. The difference in amount between a score of six and a score of eight must be equal to the difference in amount between a score of thirteen and a score of fifteen. The whole principle of comparative measures depends upon the equality of the steps all along the scale.

(3) The test should measure the achievement or ability which it purports to measure. No silent reading test meets this requirement, for example, unless the children who get the best scores in it are the ones who are the most efficient silent readers in their regular work, whether in school or out of school. No spelling test meets this requirement unless the children who get the best scores in it are those whose regular written work is most nearly free from misspelled words.

(4) The test must be brief and simple if it is intended for use by public school teachers and superintendents. To carry conviction as to its validity and importance, a test must be easily comprehended by the teacher, and the directions for giving the test and tabulating the results must not be too involved. Since the test is not an instructional device, but an instrument for measuring the results of instruction, it cannot call for any large amount of time from either the teacher or the pupils.

These criteria applied. When examined in the light of these four essential criteria the strong and weak points in the several tests described above stand out very clearly. Both the Thorndike and the Haggerty visual vocabulary tests are

almost perfectly objective, and the equality of the steps along the scale is assured by the method which was employed in deriving the tests. As to whether these tests measure only what they purport to measure, namely visual vocabulary, a question has been raised¹ but has not been answered. In reading, visual vocabulary is used not with words in isolation, but with words in their setting in sentences. How nearly a score obtained by a test using the words in isolation is a measure of the ability to recognize those same words in sentences is not known. The Jones tests are made up in both ways, one listing the words in isolation, and the other using the words in sentences, in recognition of this uncertainty. It may be doubted, too, whether a child should be expected to recognize such words as "eugene" and "jesse" when they are not capitalized. Finally, these vocabulary tests are simple and can be made brief, but the method of obtaining the class average on the record sheet is rather tedious, and will probably tend to limit materially the usefulness of the tests.

Turning to the tests for silent reading we find most of them far from perfect in objectivity. For scoring the papers in the Thorndike Scale Alpha, the Minnesota Scale Beta, and Brown's Silent Reading Test, elaborate keys are provided for the guidance of the scorer, while with Starch's tests and Gray's tests the scorer must use his judgment as to the correctness of each unit of the child's reproduction. With as large a subjective factor left in the scoring as these tests allow

¹ Jones, R. G. "Standard Vocabulary for Primary Grades"; in *Fourteenth Yearbook of the National Society for the Study of Education*, part 1, p. 43.

it is not safe to base too much upon a comparison of the results in one class with those in another, unless the papers in both cases are scored by the same person, or by persons trained for the special task. While this is true, these tests are nevertheless far more objective than the examinations in common use, and in the hands of a trained investigator their imperfect objectivity is not a serious fault.

Considered from the standpoint of what they measure these tests all present normal reading situations, and nothing in our whole educational endeavor is so important, so far as formal instruction goes, as to be able to determine how rapidly, and with what degree of comprehension, school children can read such selections as are presented. It may be questioned, however, whether a child's ability to reproduce in composition form the ideas he has gathered from the selection is not an ability quite separate from the ability to comprehend the meaning. This criticism holds but little against the Thorndike Scale, but must be considered seriously in connection with the tests of Brown, Starch, and Gray. It is to avoid this doubtful method of procedure that the Kansas Silent Reading Tests go to such a length as they do to reduce to a minimum the element of reproduction.

It may be said in this connection, however, that in avoiding much reproduction, the Kansas Silent Reading Tests have incorporated within themselves another fault, the seriousness of which has yet to be calculated. This fault is that the exercises thus made up so as to call for a minimum of reproduction partake somewhat of the nature of puzzles, and therefore do not represent normal reading difficulties.

One other consideration must be kept in mind in connec-

tion with the tests of Brown and Starch. The time involved in scoring the papers is considerable. Because of the uncertainty of uniformity of scoring in the hands of untrained persons, teachers cannot do the marking of the papers very satisfactorily, and so these tests seem to be useful mainly in special investigations where the material is to be handled by experts. The same conclusion also seems to hold with reference to Gray's Silent Reading Tests, and for the additional reason that the time required to test a class, one child at a time, is so great as to prevent general use of the tests by teachers.

The Kansas Silent Reading Tests and the Courtis Silent Reading Tests are designed to be put into the hands of the class teacher. They are very simple to administer, take but little time, and are objective. The serious question about them was pointed out above.

The Jones Vocabulary Tests are deficient in respect to the values attaching to the various words. The fact that "the" occurs 1733 times in ten primers, while "that" occurs 176 times, is scarcely a sufficient reason for scoring "the" at ten times the value of "that."

Gray's Oral Reading Test is a splendid instrument, but it is serviceable only in the hands of one carefully trained in its use. The method of testing one pupil at a time, which oral reading seems to require, makes the element of time a serious consideration at best, and this test which depends upon such elaborate marking of each paragraph read is quite burdensome. The test is objective, however, and the results obtained are very reliable measures of oral reading.

When teachers shall have acquired more generally in their

professional equipment a knowledge of standardized tests and a facility in their use, many of the above criticisms will be rendered invalid.

IV. THE SERVICE OF READING TESTS

Service to the superintendent. From the importance of reading in the general efficiency of all school work we may assume that the superintendent is vitally interested in making the instruction in reading most effective. What can reading tests reveal to him?

First, they can satisfy him and his teachers of the general status of reading in his district. It is easy for any superintendent to carry conviction among his teachers that the results in reading are not satisfactory in his district if he can show that among a group of a dozen or more neighboring cities his district stands low. The extent to which it stands low becomes a measure of the renewed earnestness needed in attacking the problem of improvement.

It is difficult for one to carry in mind a fixed standard of achievement. One gradually thinks more and more in terms of what those around him are achieving. It would have been quite impossible, for example, to convince the superintendent and teachers of the Anglo-Korean school at Songdo,¹ without a standardized test, that the children in their fifth grade could do, on the average, reading work valued at only 3.8 units, while American children who had been in school only the same number of years could score 13.2 units, or that

¹ Wasson, Alfred W. "Report of an Experiment in the Use of the Kansas Silent Reading Tests with Korean students"; in *Educational Administration and Supervision*, vol. 3, p. 98.

their sixth grade could accomplish only as much as the American third grade. It meant much to that school for its superintendent and teachers to be able to measure their school by the American standards.

As an illustration of this divergence of standards among schools of different types in this country, and among schools in different sections of the country, Table XVI is given. Test I was used in Grades III, IV, and V, in city schools, but in Grades III, IV, V, and VI in country schools because the elementary-school course is divided into nine grades in the country schools. Test II was used in the upper grades in both city and country.

The determination of status must extend beyond a general measure of reading ability. What is developed under the name of reading is in fact a complex of many abilities. This was most strikingly brought out in Cleveland, Ohio, by the recent survey which disclosed that while the city was uniformly high both in oral reading (pronouncing the successive words of a paragraph) and in speed in silent reading, it was uniformly low in the quality of reading as evidenced by the ability of the children to reproduce what they had read. Evidently the end which Cleveland was seeking to accomplish in reading was different from the end sought in the group of cities with which her work was compared. (See Fig. 23, page 297.) Be it remembered that the tests do not disclose which is right, Cleveland or the other cities, but they do disclose the difference, and in the difference lies a problem which could not have been intelligently attacked without a knowledge of the facts revealed by the tests.

Reveals wrong emphasis in teaching. Differences in the

TABLE XVI. SHOWING MEDIAN SCORES MADE WITH THE KANSAS SILENT READING TESTS

		Grade						
		III	IV	V	VI	VII	VIII	IX
First-class cities in Kansas....	Median Number of children	4.3 1873	8.8 2017	13.1 1819	13.8 1590	16.1 1546	19.7 1384	
Second-class cities in Kansas	Median Number of children	5.9 966	9.7 1067	14.3 994	14.3 1024	17.3 613	20.6 596	
Third-class cities in Kansas....	Median Number of children	4.6 373	8.2 524	11.8 471	12.5 518	14.0 352	20.6 560	
Kansas total....	Median	4.9	9.0	13.4	13.7	16.1	20.1	
Iowa total.....	Median Number of children	6.2 2371	9.5 2940	14.6 2695	14.8 2597	17.7 2143	20.6 1819	
Total from Far-Western cities.	Median Number of children	6.1 2282	10.6 2509	14.4 2643	15.0 2673	18.0 2508	20.6 2075	
Thirty-five one-room schools in Kansas....	Median	3.0	7.0	8.7	11.8	11.0	15.9	18.9
Cities in the Southern States	Median Number of children	4.7 686	8.4 723	12.3 702	11.8 602	15.4 498	19.2 350	

reading work done in the several buildings within a city may be as striking as differences among cities. In a certain Middle Western town a forceful principal of one of the ward buildings has dominated the work of the building for a good many years. The reading of the building was his particular pride. When tested for silent reading ability his children scored in

every grade but little more than half what the children in another building scored where the work was reputed to be "much less thorough." These results were made the basis of deliberations among the teachers as to the legitimate outcomes of reading, with the result that, without diminishing any one's zeal, the emphasis was transferred from oral word-pronouncing to silent thought-getting in the building where this strong principal dominates the work so effectually.

This is but one illustration of the use to be made of tests in the work of supervision. The course of study, both in its broad aspect and in its details, and the time allotment for the various phases of reading, should be made mainly in the light of measurable results. Supervision of the instruction in reading may well be done in part on the basis of what has been revealed to be the particular needs of a given room, or even of a given child. Such questions as the following may frequently be asked: —

Is meagerness of vocabulary at the bottom of this difficulty?

Is this type of reading preparing the children best for understanding their history or science work?

Are the children reading aloud those things which can best be appreciated when read aloud?

In short, the superintendent can now think in terms of standards upon such questions as the amount of attention to give to reading as a whole, and how much to each phase of reading both in his entire district and in each unit of his school system.

Service to the teacher. By far the largest service of standardized tests is being rendered to the teacher. Not only are

they enabling the teacher to check up his conception of what can justly be expected of children, but they are indelibly impressing upon his mind the absolute need for recognizing the individual differences among his pupils in respect to each problem of learning. Some child who reads well orally from his reader because he confines his study of reading very largely to the daily lesson assigned, may be doing poorly in geography or in the problems of arithmetic. These content subjects depend upon reading, but not upon the sort of oral word-pronouncing which still too largely characterizes our reading periods. This particular child needs a different sort of reading. He would be found to stand low, probably, in vocabulary; probably low in quality of silent reading. If the teacher has before him a chart upon which is recorded the standing of this child in the various aspects of reading, he will no longer assign for his study the next page or two in the reader. He needs the sort of reading which widens his vocabulary more rapidly, and centers his thought upon meaning instead of upon words. Instruction from the third reader should not be for the purpose of preparing children to read the fourth reader. Reading is a tool, and its use in the content subjects is the proper test of its efficiency. As pointed out in the report of the Survey Commission, it is probably more than a coincidence that in the Cleveland schools the per cent of failures in reading rapidly decreased from the lower to the upper grades, while failures in geography, history, arithmetic, and grammar mostly increased. Where the stress is upon oral reading and speed of silent reading rather than upon thought getting, as the tests show the case to be in Cleveland we cannot expect results in the content

subjects to be very satisfactory. Therefore a teacher, with a fairly complete diagnosis of the reading abilities of his pupils before him, cannot fail to take into account the varying needs of the individuals when directing the study of his class.

As an illustration¹ of the aid of reading tests in such diagnosis, the case of the Training School at Oshkosh, Wisconsin, may be cited. During a summer term of only six weeks pupils, by use of the Kansas Tests, the Gray Oral Test, and the Gray Silent Reading Tests, had their difficulties localized. Instruction was then given upon the points revealed to be needing attention. Twenty out of one hundred and five children were given different instruction from that given the class as a whole. Surprisingly greater results were obtained in the case of those children whose instruction was specifically adapted to their difficulties.

Service to the child. Since the beginning of schools children have been sent to school to be taught. That being the case, they wait to be told what to do, and there is the end of their responsibility. When the end of the month comes they look to their report card for a measure of their success in doing what they have been told.

A function of standardized tests, by which the child can measure his own achievements about as successfully as the teacher can, is that they bring the child into partnership with the teacher in directing the whole educative process for the child. If the child discovers by actual trial that he has

¹ Uhl, W. L. "The Use of the Results of Reading Tests as a Basis for Planning Remedial Work"; in *Elementary School Journal*, vol. 17, no. 4. (December, 1916.)

only three fourths as large a vocabulary as children of his grade the country over, or that he reads only three fourths as fast, he can be depended upon better to coöperate in overcoming the fault than when he is simply given a card every month with 70 assigned to his reading. Particularly is this true if he feels that at the end of a given period he can take his own measure again to ascertain his gain. Children should be enlisted with the teacher in the effort to select the most needful sorts of materials for their study. Where one child needs problem solving, another needs a story, while still another needs something else than reading of any kind.

Remedying the situation revealed. As results of reading tests are reported more and more widely, the conviction is gaining that we are not giving a due proportion of attention to silent reading. Children who read orally fluently are found often to master a rather meager portion of what they read. In fact it is believed that habits of reading which are established by the too exclusive use of the oral type of reading frequently work to prevent the adequate development of silent reading ability. Nevertheless teachers give but little conscious attention as yet to the development of this thought-getting ability or ability to read silently. In our schools most reading work still consists of oral expression.

Because of this growing conviction it seems proper to set down a few of the simplest hints for turning the major emphasis to silent reading. Since the Kansas Silent Reading Tests have been more widely used than others, thus establishing a more reliable standard, and since they are so very simple for the teacher to administer, results obtained

from their use will be regarded as a starting-point for these suggestions.

Suppose now that the teacher has given the test to his class. The first need is for some simple practical device by which he may make a comparison of his class with the standard scores for the same grade of pupils. To assist in doing this the table on page 81 may be used. Here are given the standard scores for each grade from the third through the twelfth. In this table are also given the twenty-five percentile (that point below which the lowest one fourth of the scores fall), and the seventy-five percentile (the point above which the highest one fourth of the scores fall). By comparing the scores of one's own class with these marks, it is easy to determine in what respect one's class is different from the standard. It should be noted here that while the score of any child as recorded in the test combines speed and accuracy, the teacher may ascertain if he chooses, by examining the paper, whether the child read rapidly and made many mistakes, or read slowly and accurately to obtain the score which is recorded.

Types of situations revealed. The situations revealed by the test fall into three types: —

First, the class may correspond closely with, or surpass the standard distribution in both median score and variability;

Second, the class may have a low median but satisfactory variability; or

Third, the class may have a satisfactory median but too wide variability.

In connection with each type of situation certain suggestions may be considered.

A normal situation. Suppose first, then, that a teacher finds his class to correspond closely with the standard distribution. The first thing to bear in mind is that these standards have been derived from the measure of actual achievement in schools of all sorts and do not necessarily represent ideal conditions of reading ability. Furthermore, silent reading has not had its proper share of attention. Certain classes and certain entire school systems have been able to achieve median scores very much higher than these standard medians and distributions with much less variability. It is a well-recognized fact, furthermore, that a very great waste in effort will continue in teaching silent reading and all the subjects dependent upon silent reading until we secure in general much less variation in ability among the members of a class than is now represented in the standard distributions. For example, the median score of the poorest one fourth of fifth-grade children is 7, while the median score of the best one fourth of fifth-grade children is 20.4. This means that while one fourth of a normal class can read one page another fourth of the class can read nearly three pages with the same degree of comprehension. Class instruction based upon common assignments of reading tasks must be, under such circumstances, most wasteful. If such assignments are well adapted to the slower pupils, they cannot at the same time bring out the best efforts of the strong ones. It behooves a teacher, therefore, to undertake to secure a distribution much less varied than the standard, at the same time he raises his median score. Suggestions for accomplishing these two things will be made in the following paragraphs.

To raise the median score. Suppose next that the test

has revealed an unsatisfactory median score, but a satisfactorily close grouping of the scores around the median, and no individuals varying strikingly from the median.

As an indication of how class medians vary among classes in the same grade Table XVII is given. From this table a teacher may know that if the median of his class, say fifth grade, falls below 11.89, his class is among the lowest one fourth of fifth-grade classes as judged by this test. In case, then, the teacher finds that his situation demands a raising of his median score the following suggestions may aid.

TABLE XVII. CLASS MEDIAN BY THE KANSAS SILENT READING TESTS

<i>Grade</i>	<i>Lowest one fourth of class medians fall below</i>	<i>Highest one fourth of class medians fall above</i>	<i>Number of classes considered</i>
3d.....	3.19	6.57	125
4th.....	7.6	11.5	136
5th.....	11.89	15.36	127
6th.....	11.78	16.33	111
7th.....	14.5	18.78	88
8th.....	17.06	22.72	78

The lowest one fourth of class medians fall below the left-hand figure, and the highest one fourth of class medians fall above the right-hand figure.

Overemphasis on oral reading. The commonest of all reasons for this situation, particularly when found in the grades below the sixth, is that the teachers have been placing chief stress upon oral reading. Where children are required to give their attention mainly to the correct pronunciation of words, the correct enunciation of sounds, and the correct inflection of the voice in passing over the several punctuation marks, not much growth in the power to comprehend

meaning in the language can be expected. Where the children study their reading lesson with the point of view of being able to respond in this way, they fasten upon themselves the habit of watching for words whose pronunciation they are not sure of, or they form the habit of reproducing the sounds of syllables, thus establishing the practice of moving the lips and other speech organs when reading silently. Frequently both these habits fix themselves upon children whose reading is judged mainly by the daily oral performance. When either or both habits become fixed a real struggle is required to break them. Unless they are broken, however, the child suffers a severe handicap the rest of his reading life. Many men and women of mature years are still paying the price of those habits fixed in youth. They read but little faster silently than they can pronounce the words orally, because their speech organs make all the motions of the successive words as the reading proceeds.

Care from the beginning. To be on guard against these two habits care must be exercised from the very beginning. Children in the primary grades should have exercises from the start in which the meaning is the only significant element, and the response is not in terms of words said, but things done, or interpretations made. For example, let it be the usual thing for the child to carry out the directions contained in the word or sentence. The primary teacher should be supplied with some hundreds of cards upon which such sentences or short paragraphs as the following are printed or written:—

- (1) Draw a picture of a flag on the blackboard.
- (2) Make a sound like a cross kitty makes when a dog chases her.

- (3) Hide behind the door.
- (4) Play that you are carrying a cup full of water and do not wish to spill any of it.

These cards should be graded in such a way that certain ones will contain only the words taught in the first reading lessons. As more words are learned, more cards will become available. Word drills should then divide time very generously with these practice cards which emphasize attention to meaning.

Variety in handling the exercises may be introduced in scores of ways which will readily occur to a resourceful primary teacher. Many other devices having the same aim will also occur to the teacher. The essential thing is that practice in translating written or printed language into action instead of words should be started early, thus producing the habit of advancing through a paragraph by thought-units rather than by letters, syllables, or words.

Reading above the primary grades. In grades above the primary the problem is fundamentally the same as stated for the primary, but the devices must vary.

First, whenever reading is done orally, be sure that what the child is reading is new to most of his listeners. Be sure, too, that the other pupils are listening, and not following along with the reader in another copy of the same book. No method of reading is more faulty in intermediate grades than that in which other members of the class are watching for a word error of the reader, ready to call attention at once to such a mechanical mistake. This method centers the attention of the reader constantly upon the mechanics and never develops the habit of attending first to the thought. Whereas,

if the reader realizes that his hearers know nothing of the content of his selection except what they gather from his reading, then giving the thought instead of pronouncing the words becomes the controlling factor in his consciousness. It follows from this that only selections, the thoughts in which are vital to children, should be used as subject-matter for such reading. Then let the one who has read such a selection defend the selection against questions or criticisms of the class. In short, center attention upon the meaning, even at the expense, if necessary, of accuracy in pronunciation, enunciation, and expression.

Second, let the amount of reading which is compellingly interesting be increased. Supplementary reading in geography, history, science, and literature should be given a larger place. Require that the reports made upon such readings be rather exact, but let the selections be reasonably easy for the children. Gain in facility in silent reading cannot be secured by holding the children to selections which are so difficult that word-troubles absorb all the attention. One must be able to go with ease through the successive thoughts before the habit of attending to the thought can be acquired.

Third, make all the industrial and playground exercises give a far greater measure of service in teaching reading than they now commonly give. How singularly shortsighted we are to ask a child to follow the directions printed in his arithmetic for finding the per cent that one number is of another, but employ a teacher to give orally the directions for playing a new game, making a raffia basket, or planting beans. The very things which come nearest the

natural interests of the children, concerning which they would most zealously read if they had the paragraphs containing the needed directions, are given to them orally. When interesting school exercises require a careful following of directions, then those directions make the most effective silent reading material. But in practice we seldom make use of them. This fault is due to a failure to understand the distinction between the aim of the intermediate grades and the aim of the upper grades. If we realized that all the work of the intermediate grades should be made to develop skill in using the tools of learning, then we should not conduct these exercises without making them aid in teaching reading.

Reading in the upper grades. Passing now to the situation presented when the score of a class above intermediate grades is found to be low, we have the most serious task of all. The junior high-school or upper-grade pupil should be able to proceed with his school tasks without much attention to the tools he is using. It is not the primary function of this department of the school system to increase the children's facility in the handling of these tools. However, success in nearly all the tasks undertaken in the upper grades depends upon the skill which the children are expected to possess in the tool subjects. A compromise is, therefore, necessary, if children in the junior high school, or seventh and eighth grades, are found deficient in their ability to read silently. A few suggestions are here offered in the hope that some help may come from them, although it is realized that correcting reading faults at this stage is very difficult.

First of all, the children's own conscious efforts should be

obtained in the direction of correcting the faults. Then, too, the teacher should see that he is observing the same fundamental principles stated for the intermediate grades. Comprehension, and not mechanics, must be made the test of all reading, whether in history, science, or literature. The material selected for use must be sufficiently easy so that the children are not tied up in word or language difficulties. Again, to overcome the habit of proceeding by too small units, practice must be afforded in advancing by short sentences or phrases.

In case the trouble seems to be that the children read fluently enough orally, but get little of the thought, introduce a great deal of the sort of reading requiring close attention to the thought. For example, use rule books for football, basketball, and the like for those interested in games; catalog descriptions; directions for making certain stitches; the more involved arithmetic problems, and so on. These things possess a minimum of word difficulty and a maximum of thought difficulty. They require the imagination to construct a picture little by little and hold it up for constant modification as the reading proceeds. Thus, attention is focused on thought.

Where the class appears to have the right habits of reading silently but have had insufficient practice, the obvious suggestion is to give them all the practice possible. Much supplementary reading upon which they make only meager reports, if any, will help. Try to secure as much general home reading as possible. See that an abundance of interesting things is available for reading and stimulate interest by having the children's criticisms of them given before the class.

Where variability is too wide. We come now to the situation where the range of ability within a class group is too wide. Here the problem is a different one from that presented by a class with a low average ability. Here different treatment is needed by the different members of the class. It is in this situation that a teacher needs to keep the diagnosis of the abilities of his class graphically before him. Each part of the day's work must call for exercise from each child in those particular skills most needing development. No other service of scientific measurement in education promises more for the future than just this.

Uniformity in instruction for all the members of a class widens variability among them, making the weak ones relatively weaker and the strong ones relatively stronger. To prevent this widening variability more attention must be given to individual instruction. This does not mean a leveling of all members of a class, but rather affording a maximum of opportunity to each member to do those things most needful to him. Those things which he can already do well he is not required to do, even though some other members of the class need them.

Those children falling far below the median of the class should be given special physical examination to discover if possible the cause. Sometimes eyesight is found to be poor. Frequently some other physical defect has prevented normal mental growth. Sometimes an examination by means of approved intelligence tests, such as the Binet-Simon tests,¹

¹ See especially *The Measurement of Intelligence*, by L. M. Terman. (Houghton Mifflin Company, Boston, 1916.) A simple guide for the use of the intelligence scale.

reveals that the child is mentally incapable of doing work of the regular school type.

The difficult but normal case; suggestions for helping. If, however, the child is found nearly normal physically and mentally, but has not developed ability to get meaning from printed language, he presents a problem in instruction calling for the best professional skill to solve. The following suggestions may help: —

First, make use of the devices suggested for raising the general average of ability in silent reading given in earlier pages. They can be used with the pupils who are below standard, while the other members of the class have other assignments.

Second, it is quite certain that a pupil far below the median in this basic ability has never made use of printed language to secure help in satisfying his own childish desires. If possible, situations must be brought about in which his desires or plans depend for their fulfillment upon his reading. It may be, for example, that his mother or father has been in the habit of reading stories to him. If so, and he can be made to be keenly interested in a story by having a part of it read to him, he should have to read the rest himself to satisfy his desire to know the rest of the story. Possibly he would like to be the leader in an occasional nature study excursion, but, of course, it will be expected that he look up information concerning the things they see on the trip and be able to report later to the group. That is the business of the leader. Or, he might umpire the baseball game if he made sure of the rules; or assign the parts in the coming school entertainment, if he read the various parts carefully so as to be able

to make a wise assignment; or score the class compositions on the basis of which was most interesting. Such a list of possible opportunities for calling into service a child's silent reading ability might be largely extended. The two things to guard against are (1) making reading a punishment and (2) confusing child need with school need. The thing to be accomplished is to give the child a chance to do something which he really wishes to do but cannot do without reading.

Third, see to it that the regular work assigned to him in the school subject is not too difficult. Skill in silent reading is developed in early years by reading widely in relatively easy matter rather than from reading intensively a very small volume. The very slow reader is usually one who has never caught the knack of disregarding words and attending to thought. In order to acquire this knack it is necessary to have easy reading. Therefore, while the class is studying a text in history, let the slow reader be assigned some early biographical story bearing upon the events studied.

Let it be said in conclusion that the chief aim of these suggestions shall have been attained if they serve to advance the conviction that there is a real problem of teaching silent reading, distinct from the problem of teaching oral reading. If the conviction once gets hold of the teachers, the problem is half solved.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What are the chief methods by which adults add new words to their vocabularies? Are more new words learned from the context in which they appear, or from the dictionary? What can you say concerning the best way to increase the vocabulary of children?

2. What are some of the other factors besides vocabulary involved in silent reading? In what grades is vocabulary the most important factor? Make some suggestions for guaranteeing the intimate association of the mental concept which a word symbolizes, and the word itself when it is encountered in word drills.
3. What is the significance of speed in reading? Is there any truth in the rather common belief that one who reads slowly "gets more out of what he reads"? If you do not know the answer, can you devise some way to test it out in your class? Compare your own silent reading rate with that of some equally well educated friends.
4. What are the chief dangers involved in having much oral reading in the lower grades? Can these dangers be safeguarded? What types of reading matter do you now read orally outside the schoolroom? Are these the types which your pupils are asked to read orally?
5. What are the circumstances under which you last read aloud? Do your pupils have the same incentives for reading clearly and interestingly that you had on that occasion?
6. What are some of the things you do to assist your pupils in developing ability to comprehend the meaning of the printed page? Do you know of faulty habits which some of them have which prevent their centering attention upon the meaning? Do you know which pupils read with accuracy? Which with speed?
7. Can you think of a more simple way than Thorndike used to test whether a child is familiar with a given word or not? Pick out about six pages of material from unfamiliar texts in various subjects, have the pupils read them, lightly underlining all the words which they do not know. Rank the children according to their vocabulary knowledge. Now use some standardized vocabulary test and see how the two rankings compare. What more does the standardized test accomplish even if the rankings are practically the same?
8. How long does it take you to become familiar with the reading difficulties of each child when you receive a new class of, say, thirty children? Would you consider it economical if some tests were available by means of which you could discover these difficulties as well as others the first day and thus prepare a chart of each child's instructional needs? How long at the beginning of a term could you afford to spend in making such a diagnosis?
9. Do you think the abilities of your class in silent reading vary as widely as is indicated by the table of scores for the Kansas Silent Reading Tests? What do you do to take into account the variability which you recognize does exist?
10. Think of the last examination you gave in reading. How satisfactory do you think it was from the standpoint of objectivity? Of the equality of the questions? Did it test satisfactorily what you are striving to teach in reading?

11. What are the advantages of standardized tests in reading as listed in this chapter over the ordinary tests in reading? Name the tests here discussed, and briefly describe each.
12. Which test do you think will give you the most helpful information? Why?

BIBLIOGRAPHY

Only the most important references are given here. For other references see footnotes in the chapter.

I. SILENT READING

1. *The Kansas Silent Reading Tests*, devised by F. J. Kelly. For copies of the tests, address Bureau of Educational Measurements and Standards, Emporia, Kansas. Test I is for Grades III, IV, and V; Test II, for Grades VI, VII, and VIII; Test III, for Grades IX, X, XI, and XII.

REFERENCE: Kelly, F. J. "The Kansas Silent Reading Tests"; in *Journal of Educational Psychology*, February, 1916.

2. *Gray's Silent Reading Tests*. Copies of the tests may be obtained from William S. Gray, School of Education, University of Chicago, Chicago, Illinois.

REFERENCES: Judd, Charles H. *Measuring the Work of the Public Schools*. (Cleveland Foundation Survey Report, Cleveland, Ohio.) Also for sale by The Russell Sage Foundation, New York City.

Gray, William S. "A Study of the Emphasis on Various Phases of Reading Instruction in Two Cities"; in *Elementary School Journal*, vol. 17, pp. 178-86. (November, 1916.)

Gray, William S. "A Coöperative Study of Reading in Eleven Cities in Northern Illinois"; in *Elementary School Journal*, vol. 17, pp. 250-65. (December, 1916.)

Gray, William S. *Studies of Elementary School Reading through Standardized Tests*. (Supplementary Educational Monographs, University of Chicago Press.)

3. *Brown's Silent Reading Test*. Copies may be obtained from H. A. Brown, Bureau of Research, 25 Capitol St., Concord, New Hampshire.

REFERENCES: Brown, H. A. *The Measurement of the Ability to Read*. (Bulletin no. 1, Bureau of Research, Concord, New Hampshire.)

Brown, H. A. "The Measurement of the Efficiency of Instruction in Reading"; in *Elementary School Teacher*, vol. 14, pp. 477-90. (June, 1914.)

4. *Starch's Silent Reading Tests*. Copies may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.

REFERENCE: Starch, Daniel. "The Measurement of Efficiency in Reading"; in *Journal of Educational Psychology*, January, 1915.

5. *Thorndike's Scale Alpha for Measuring the Understanding of Sentences*. Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCES: Thorndike, E. L. "An Improved Scale for Measuring Ability in Reading"; in *Teachers College Record*, November, 1915, and January, 1916.

Haggerty, M. E. *The Ability to Read: Its Measurement and Some Factors Conditioning it*. (Indiana University Studies, no. 34.)

6. *The Minnesota Scale Beta*. Copies may be obtained from the Bureau of Coöperative Research, University of Minnesota, Minneapolis, Minn.
7. *Fordyce's Scale for Measuring the Achievements in Reading*. Copies may be obtained from the University Publishing Company, Lincoln, Nebraska.
8. *Courtis's Standard Research Tests in Silent Reading*. Copies may be obtained from S. A. Courtis, 82 Eliot Street, Detroit, Michigan.

II. ORAL READING

1. *Gray's Oral Reading Test*. Copies may be obtained from William S. Gray, School of Education, University of Chicago, Chicago, Illinois.

REFERENCES: Judd, Charles H. *Measuring the Work of the Public Schools*. (Cleveland Foundation Survey Report, Cleveland, Ohio.)

Gray, William S. *Studies of Elementary School Reading through Standardized Tests*. (Supplementary Educational Monographs, University of Chicago Press.)

III. VOCABULARY TESTS

1. *Jones's Scale for Teaching and Testing Elementary Reading*. Copies may be obtained from the Rockford Printing Company, Rockford, Illinois.

REFERENCE: Jones, R. G. "Standard Vocabulary"; in *Fourteenth Yearbook of the National Society for the Study of Education*, part 1, pp. 37-42.

2. *Haggerty's Visual Vocabulary Test for Grades 1 and 2*. Copies may be obtained from Bureau of Coöperative Research, University of Minnesota, Minneapolis, Minn.

REFERENCE: Haggerty, M. E. "Scales for Reading Vocabulary of Primary Children"; in *Elementary School Journal*, vol. 17, pp. 106-15. (October, 1916.)

3. *Thorndike's Visual Vocabulary Scale Alpha*. Copies of the scale may be obtained from Bureau of Publications, Teachers College, New York City.

REFERENCES: Thorndike, E. L. "The Measurement of Ability to Read"; in *Teachers College Record*, September, 1914.

Thorndike, E. L. "The Measurement of Achievement in Reading; Word Knowledge"; in *Teachers College Record*, vol. 17, pp. 480-54. (November, 1916.)

4. *Starch's English Vocabulary Test.* Copies of this test may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.

IV. GENERAL REFERENCES

Anderson, Homer Willard. *Measuring Primary Reading in the Dubuque Schools.* (The Harris-Anderson Tests, Dubuque, Iowa, 1916.)

Gray, William S. "Methods of Testing Reading"; in *Elementary School Journal*, vol. 16, pp. 281-98. (February, 1916.)

Otis, Arthur S. "Considerations Concerning the Making of a Scale for the Measurement of Reading Ability"; in *Pedagogical Seminary*, vol. 23, pp. 528-49. (December, 1916.)

Pinter, Rudolph, and Gilliland, A. R. "Oral and Silent Reading"; in *Journal of Educational Psychology*, vol. 7, pp. 201-12. (April, 1916.)

Richards, Alva M., and Davidson, Percy E. "Correlations of Single Measures in Some Representative Reading Tests"; in *School and Society*, vol. iv, pp. 375-77. (September 2, 1916.)

Uhl, W. L. "The Use of the Results of Reading Tests as Bases for Planning Remedial Work"; in *Elementary School Journal*, vol. 17, pp. 266-75. (December, 1916.)

Ziedler, Richard. "Tests in Silent Reading in the Rural Schools of Santa Clara County, California"; in *Elementary School Journal*, vol. 18, pp. 55-62. (September, 1916.)

CHAPTER IV

SPELLING

I. THE PROBLEM OF MEASUREMENT IN SPELLING

Difficulties encountered. The measurement of spelling ability involves certain difficulties which are peculiar to the subject. Shall spelling ability be construed to mean ability to spell words when the attention is focused upon the ideas which are being expressed in writing, rather than upon the spelling of the words? Or is it the ability to spell words when the attention is focused upon the spelling of words, as in the case of dictated spelling lists? Upon the basis of what words shall one's spelling ability be measured? Shall spelling ability be defined in terms of the per cent of correct spellings of a limited group of frequently used words, or shall it be defined in terms of the extent of one's spelling vocabulary.¹

It appears to be the consensus of opinion that one needs to be able to spell correctly the words used frequently, with a minimum of attention or automatically, as is the case in writing letters, compositions, school exercises, and the like. Otherwise, it is impossible for one to focus his attention upon the ideas which are being expressed. This type of spelling ability is defined as the per cent of words spelled

¹ There are other phases of spelling ability, such as an ideal of correct spelling which will insure the use of the dictionary in the case of words concerning whose spelling a writer is uncertain, or the ability to detect words which are misspelled. However, instruments for measuring these phases of spelling ability have not been devised.

Fig. 5. MEASURING SCALE FOR
ABILITY IN SPELLING

Russell Sage Foundation, New York City

Division of Education

Leonard P. Ayres, Director

All the words in each column are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top indicate about what per cent of correct spellings may be expected among the children of the different grades. For example, if 20 words from column H are given as a spelling test it may be expected that the average score for an entire second grade spelling them will be about 79 per cent. For a third grade it should be about 92 per cent, for a fourth grade about 98 per cent, and for a fifth grade about 100 per cent.

The limits of the groups are as follows: 50 means from 46 through 54 per cent; 58 means from 55 through 62 per cent; 66 means from 63 through 69 per cent; 73 means from 70 through 76 per cent; 79 means from 77 through 81 per cent; 84 means from 82 through 86 per cent; 88 means from 87 through 90 per cent; 92 means from 91 through 93 per cent; 94 means 94 and 95 per cent; 96 means 96 and 97 per cent; while 98, 99 and 100 per cent are separate groups.

By means of these groupings a child's spelling ability may be located in terms of grades. Thus if a child were given a 20 word spelling test from the words of column O and spelled 15 words, or 75 per cent of them, correctly it would be proper to say that he showed fourth grade spelling ability. If he spelled correctly 17 words, or 85 per cent, he would show fifth grade ability, and so on.

The data of this scale are computed from an aggregate of 1,400,000 spellings by 70,000 children in 84 cities throughout the country. The words are 1,000 in number and the list is the product of combining different studies with the object of identifying the 1,000 commonest words in English writing. Copies of this scale may be obtained for five cents piece. Copies of the monograph describing the investigations which produced it may be obtained for 30 cents each, including the scale. Address the Russell Sage Foundation, Division of Education, 130 East 22d Street, New York City.

correctly. In addition, it is desirable that one should be able to spell a number of words which are used only occasionally. In the case of the more difficult and unusual of these words it is probably sufficient if one is able to spell them correctly when attending to them.

We shall consider first what the words most commonly used are, and how to measure the ability of pupils to spell them.

The foundation words of the English language. In determining the most commonly used words the method employed has been to examine written material of several types, such as letters, newspapers, and children's compositions, and to obtain a list of the words used and the number of times each word occurs. Ayres¹ has combined the results of four such studies. Two of these studies were based on letters, the third upon newspapers, and the fourth upon selections of standard literature. The material examined in the four studies aggregated 368,000 words, written by twenty-five hundred different persons.

It was the original intention of Ayres to identify the two thousand most commonly used words, but this was impossible because the material examined was found to consist of a few words used many times, and of a larger number of words used only a very few times. It was found that fifty different words were used so frequently that they made up approximately half of the material examined. In order to secure a list of the thousand most frequently used words it was necessary to include words which were found only

¹ Ayres, L. P. *Measurement of Ability in Spelling*. (Bulletin of the Division of Education, Russell Sage Foundation, New York City, 1915.)

forty-four times in the 368,000 words of material examined. This list of one thousand words is the best statement which we have of the words that form the core or foundation of the English language.

Another important study has been made by Jones.¹ He collected themes from pupils in grades two to eight inclusive. In order that a record of the complete writing vocabulary of each pupil might be obtained, a large number of compositions were written, the number per pupil ranging from 56 to 105. A total of 75,000 themes, consisting of a total of 15,000,000 words and written by 1050 pupils residing in four States, were examined. However, only 4532 different words were used by these pupils.

Making a spelling test. After we have a list of the most commonly used words, such as Ayres has given us, there remains the problem of constructing a test to measure the automatic spelling ability of pupils. It is a well-known fact that some words are more difficult to spell than others.² The words included in a test either must be equal in difficulty, or their relative difficulties must be known. Otherwise we will be using a measuring instrument consisting of unequal units, but will be considering the units to be equal. This condition would cause the measures made with such an instrument to be inaccurate. Pupils spelling only a few words correctly would receive a score higher than they

¹ Jones, N. Franklin. *Concrete Investigations of the Material of English Spelling.* (University of South Dakota Bulletin. 1913.)

² The spelling difficulty of a word has two interpretations. It may be taken to mean the difficulty which children experience in learning to spell it. It may also refer to the frequency with which it is misspelled. The latter meaning will be used in this chapter.

deserved, and the bright pupils would receive a score lower than they deserved.

The spelling difficulty of words for a given group of children may be determined by having the words spelled by them. From the per cent of correct spellings of each word the relative difficulty of the words may be calculated.¹ Words which are misspelled an equal per cent of times by pupils of a given grade are equal in difficulty for that group. In the absence of this information it is practically impossible for a teacher to judge the difficulty of the words. Buckingham concluded that the judgment of a single teacher is almost of no value. "It may be good and it may be bad; and it is about as likely to be the one as the other."

II. SPELLING SCALES

1. *The Ayres Spelling Scale*

How constructed. To determine the words of equal difficulty and the relative difficulty of the groups of words, Ayres divided the thousand words into fifty lists of twenty words each. Each list of words was spelled by the children of two consecutive grades in a number of cities. The thousand words were then divided into another fifty lists of twenty words each. Each of the new lists was spelled by the children in four consecutive grades. In all 70,000 children spelled twenty words, making a total of 1,400,000 spellings, or an average of fourteen hundred spellings of each of the thousand words.

¹ See Buckingham, B. R. *Spelling Ability, Its Measurement and Distribution*, p. 35, and following for the method. (*Teachers College Contributions to Education*, no. 59, 1913.)

Upon the basis of this information Ayres classified the words into twenty-six groups, the words of each group being approximately equally difficult for school children of a given grade.¹ This classified list, together with the per cent of pupils in each of the grades who spelled the words of each list correctly, has been printed with the title, *Measuring Scale for Ability in Spelling*. This scale is reproduced as Fig. 5.²

Pupils who are not tested. When a pupil spells correctly all of the words of a given list, we do not have a measure of his spelling ability. We simply know that he can spell these words correctly; we do not have any information concerning how far beyond this list his spelling ability extends. In fact, the pupil has been given no opportunity to show how well he can spell. It is a well-known fact that the pupils of any grade or of any class are not equal in ability, but exhibit a wide range of ability. Thus in testing a class it is necessary to use words for which the average per cent of correct spellings is less than one hundred. Ayres recommends that in making a test for the pupils of a given grade the words be taken from the column for which an average of eighty-four per cent of correct spellings may be expected.³

Figure 6 represents a typical result of using the words

¹ For the details of the method employed see Ayres, L. P. *Measurement of Spelling Ability*, pp. 22-35.

² The author is indebted to Dr. Ayres for permission to reproduce this scale.

³ The reader should not confuse scores or measures of ability with school marks. The per cent of correct spellings is a measure. The school mark is the meaning which the school attaches to that measure. The fact that both the measure and the school mark may be expressed in per cents does not make them the same. See chapter ix for a more complete discussion of this point.

chosen as Ayres recommends. The class average is eighty-four per cent, but those pupils who spelled all of the words

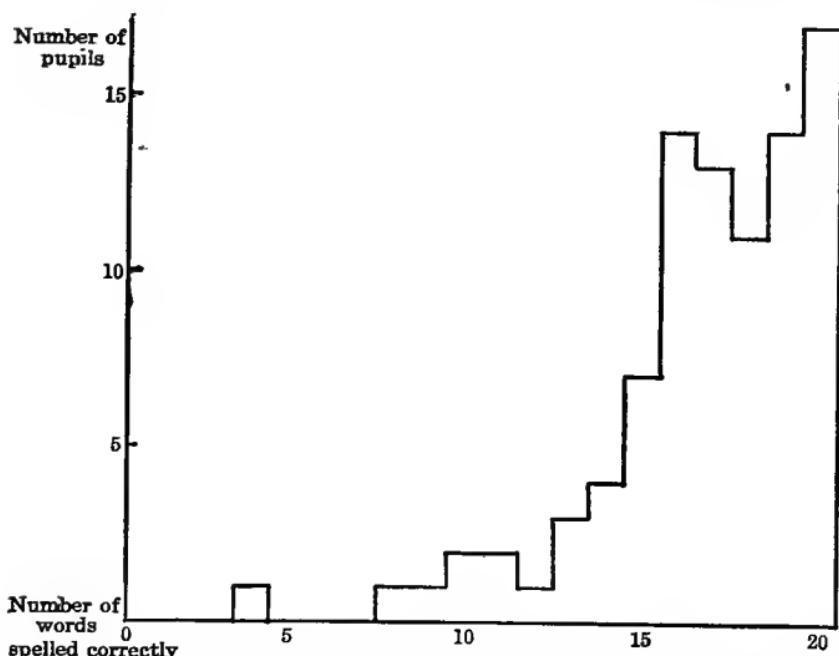


FIG. 6. SHOWING THE DISTRIBUTION OF 91 PUPILS ACCORDING TO THE NUMBER OF WORDS SPELLED CORRECTLY

Class average, 84 per cent.

correctly have not been tested. Those who misspelled only one or two words probably have not been tested satisfactorily.

Otis¹ presents facts from which he concludes that the most reliable measures of spelling ability are obtained by using words for which there is an average of fifty per cent of correct spellings. In support of this conclusion he points out that a list of words for which the average per cent of correct

¹ Otis, A. S. "The Reliability of Spelling Scales"; in *School and Society*, vol. 4, p. 753.

spellings was either zero per cent or one hundred per cent would yield a measure of zero reliability. Likewise a list of words for which the average per cent of correct spellings was ten per cent or ninety per cent would yield measures only slightly more reliable. Hence it seems natural that the most reliable measures would be obtained by using a list for which the average per cent of correct spellings was fifty. On the other hand, some writers claim that it is not wise to have pupils spell words incorrectly. Every repetition tends to fix a habit.

Ayres gives no satisfactory justification for recommending the choice of words for which an average of eighty-four per cent of correct spellings may be expected. When measuring the spelling ability of children in Springfield, Illinois, Ayres used words for which seventy per cent of correct spellings had been obtained. For the Survey of Cleveland, Ohio, the words were chosen from columns for which the average per cent of correct spellings was seventy-three. Thorndike has used words for which the per cent of correct spellings is fifty.¹ For these reasons it is probably best to choose words from columns for which the average per cent of correct spellings is approximately seventy.

How many words to use. Another question which must be considered in making a spelling test is the number of words it is necessary to use. In general the ability to spell one word is separate and distinct from the ability to spell any other word. Ability to spell, therefore, consists of a large number of abilities to spell specific words. This being the case it

¹ Thorndike, E. L. "Means of Measuring School Achievement in Spelling"; in *Educational Administration and Supervision*, vol. 1, p. 306.

would be necessary to use all of the thousand words of Ayres's list in order to obtain a complete and accurate measure of a pupil's ability to spell the most commonly used words. However, it is possible to secure a measure which is representative of the pupil's ability to spell these words by using a smaller number of words. This is possible in just the same way that it is possible to determine the quality of a load of wheat or a vat of cream by the examination of a sample.

How many words are necessary in making a spelling test depends upon what is desired. Relying upon the theory of random sampling, Thorndike believes a small number of words is sufficient to measure the spelling achievement of a large school system. A test consisting of only ten words has been used in a number of school surveys. This number is probably sufficient for the measure of a large school system, but if it is desired to obtain a measure of the spelling ability of individual pupils, a larger number must be used. Otis¹ says that a twenty-five word test gives a very poor measure of individual ability, and that at least one hundred words should be used, better four hundred or five hundred words. Starch recommends the use of two hundred words.

Methods of giving the test. The words which make up a test may be dictated to the pupils as separate words, or they may be embedded in sentences which are dictated. Furthermore, the dictation of the sentences may be timed so that the pupils are forced to write at their normal rate. Investigation has shown that the per cent of correct spell-

¹ *Loc. cit.*, pp. 679, 682.

ings is higher when the words are dictated separately than when they are dictated in timed sentences and the pupils are forced to write at their normal rate. According to Courtis the per cent of correct spellings is about five greater when the words are dictated in lists. Fordyce has found this difference to be between ten and fifteen per cent.

This means that the ability to spell words when the attention is focused upon the spelling, as is the case when the words are dictated separately, is not the same as the ability to spell the same words when the act of spelling is in the margin of one's attention. In writing letters, compositions, and the like, the spelling must be carried on in the margin of the attention because the ideas which are being expressed must occupy the focus of the attention. This is particularly true of the foundation words of the language such as we have in the Ayres list. The words of this list constitute over ninety per cent of the words we use. Hence, by using the words embedded in sentences and dictated rapidly enough to force the child to write at his normal rate, we measure the spelling ability which functions in one's everyday writing.

Letters per minute. Pupils may be caused to write at approximately their normal rate by dictating the sentences at that rate. The standards for speed of handwriting are as follows in terms of letters per minute: second grade, 36 letters; third grade, 48 letters; fourth grade, 56 letters; fifth grade, 65 letters; sixth grade, 72 letters, seventh grade, 80 letters; eighth grade, 90 letters. The dictation of a sentence requires some additional time, probably 10 per cent. For example, in the case of the sixth grade, instead of dictating at the rate of 72 letters in one minute, 66 seconds should be

allowed for words totaling 72 letters. The number of seconds to be allowed per letter for the several grades are as follows:—

<i>Grade</i>	<i>Seconds per letter</i>
II.....	1.83
III.....	1.38
IV.....	1.18
V.....	1.01
VI.....	.92
VII.....	.83
VIII.....	.73

If the sentences contain more than thirty to forty letters, they should be dictated in sections, so that the pupils' writing will not be slowed up by trying to recall what has been dictated. Furthermore, tests of speed in handwriting have showed that all pupils do not normally write at the same rate. For this reason provision must be made for those pupils who are accustomed to write more slowly than the standard rate. This can be done by having none of the test words come at the end of the sentences, and requiring all pupils to begin upon the next sentence as soon as it is dictated, even if they have not finished writing the preceding.

What the Ayres Scale really is. Strictly speaking, the Ayres "Measuring Scale for Ability in Spelling" is not a measuring instrument in itself, but rather a list of the foundation words of the English language, classified into twenty-six groups according to spelling difficulty. The teacher should use this list as a source of words for constructing spelling tests. These tests should be constructed according to the following principles, which have been considered in the preceding pages:—

1. The words for a test should be chosen from one column,

so that these will be equally difficult to spell. If this is not done, the inequality of difficulty must be recognized, if an accurate measure is secured.

2. Twenty words are probably sufficient to secure a reliable measure of the spelling ability of a class. At least fifty words should be used to secure a reliable measure of the spelling ability of individual pupils. More accurate measures will be obtained by using one hundred words. In the case of the upper grades it will be necessary to use words from more than one column. When this is done the relative difficulty of the words must be recognized to secure an accurate measure.

3. In order that the words may be difficult enough to really measure the spelling ability of all pupils the words should be chosen from columns for which the standard per cent of correct spellings is approximately seventy. For the lower grades it is probably best to use words for which the standard per cent of correct spellings is from fifty to sixty-six. If the words are to be used in timed sentences it will probably be satisfactory to use easier words.

4. The words should be embedded in sentences, and the sentences dictated at approximately the standard rate of handwriting for the grade. Test words should not occur at the end of the sentences.

Directions for giving a timed sentence test. The following test has been constructed in accordance with the above principles. The directions given below should be followed in giving it:—

1. See that the pupils are provided with two or three sheets of paper, and with either pencil or pen and ink. If

pencils are to be used, they should be well sharpened. If pen and ink are used, good pen points should be provided.

2. Say to the pupils: "I have some sentences which I want you to write as I dictate them. I am going to dictate them rather rapidly, possibly more rapidly than some of you can write. If you have not finished writing one sentence when I begin to dictate another, I want you to leave it and begin on the new sentence. If there are any words you cannot spell you may omit them. Take time to dot your *i*'s and cross your *t*'s. If you have any question about what you are to do, ask it now because you cannot ask questions after I begin to dictate."

3. Make certain that all pupils understand what they are to do. It is well to give a short preliminary practice in writing from dictation if the pupils are not accustomed to it. For this purpose use some simple selection.

4. Dictate the first sentence when the second hand of your watch is at 60. When it reaches 27, dictate the second sentence. When it reaches 13, dictate the third, and so on. Dictate the sentences distinctly, but do not repeat. It is advisable for the teacher to practice dictating the sentences according to the directions before attempting it with a class.

5. Stop the pupils promptly at the time indicated. Allow no corrections to be made. Ask the pupils to turn their papers over and write their name and grade. Collect the papers.

**A TIMED SENTENCE SPELLING TEST OF FIFTY WORDS TAKEN
FROM COLUMN O¹**

(Arranged for a Fourth Grade)

- (60) The *public appear* to want it.
- (27) The *population* of the *district* is five hundred.
- (13) We *refuse* to *attend* the meeting.
- (44) My *uncle* will *remain until* the man comes.
- (23) *Forty* members of the *fourth company* will drill.
- (9) The *judge knew* the *chief of police* was there.
- (52) The *whole address* was tiresome.
- (23) The *comfort* of our *friend* is to be considered.
- (7) A *perfect figure* was drawn.
- (33) During the month of *August* we *elect* a teacher.
- (17) The *navy* will go *farther* away.
- (45) A *board* from the *shed* is needed.
- (15) The house was *between* the *jail* and the store.
- (58) I am *getting* the *second* order.
- (26) The *station* is *worth* more money.
- (57) *Madam* will return *Thursday* morning.
- (32) We *don't request* attendance.
- (60) What is the *objection* to a *personal* letter?
- (41) A *sudden change* in the *weather* will come soon.
- (25) *Duty before pleasure* is an old saying.
- (2) *Eight men intend* to retire from the house.
- (42) It would be *proper* to call.

When the second hand reaches 7, stop the work.

2. *The Buckingham Spelling Scale*

Starting with a list of about five thousand words common to at least two out of five spelling books, Buckingham² selected by means of an elaborate statistical procedure two

¹ This test is one of a series devised by the author. Specimen copies may be obtained from the Bureau of Educational Measurements and Standards, Emporia, Kansas.

² Buckingham, B. R. *Spelling Ability, Its Measurement and Distribution.* (Teachers College Contributions to Education, no. 59, 1913.)

lists of twenty-five words each. The purpose of the selection was to secure "words which were easy enough in the third grade and hard enough in the eighth grade to afford a test in those and therefore intermediate grades, and which showed regular increases in per cent correct from grade to grade." The difficulty of each word was determined in terms of a common unit. Since the difficulty of each word is known the entire list may be used as a test, or any desired number may be selected from it.¹ However, the fact that the scale consists of only fifty words limits its usefulness to the general measurement of groups of pupils.²

3. Starch's Spelling Scale

Measuring the extent of ability to spell. In securing a measure of the number of words which a pupil or a class can spell correctly we are not concerned simply with the most commonly used words of the English language, but rather with all the words of the language.³ Starch has prepared a set of six word lists, each consisting of one hundred words. The words for these tests were chosen by taking the first defined word on the even-numbered pages in Webster's New International Dictionary (1910 edition). Technical, scientific, and obsolete words were discarded from the list. The re-

¹ See E. L. Thorndike, "Means of Measuring School Achievement in Spelling" (*Educational Administration and Supervision*, vol. I, p. 306), for an arrangement of words from Buckingham's list in sentences.

² See on this point Tidyman, W. F., "A Descriptive and Critical Study of Buckingham's Investigation of Spelling Efficiency"; in *Educational Administration and Supervision*, vol. II, pp. 290-304.

³ Starch, Daniel. "The Measurement of Efficiency in Spelling, and the Overlapping of Grades in Combined Measurements of Reading, Writing, and Spelling"; in *Journal of Educational Psychology*, vol. 6. (March, 1915.)

maining six hundred words were arranged alphabetically according to size. When arranged in this way they were divided into six lists of one hundred words each by taking the first, seventh, thirteenth, and so forth for the first list; the second, eighth, fourteenth, for the second list; and so forth. Each of these lists consists, therefore, of one hundred words taken at random from the non-technical words of the English language. Such a list is an instrument for measuring the extent of a pupil's correct spelling vocabulary.

The words in each list are arranged according to the number of letters they contain. Ayres¹ found for the words of his list a very high correlation between the length of words and spelling difficulty. Assuming this to be true for Starch's list, the words may be considered to be arranged in the general order of spelling difficulty by groups. A pupil's score is the number of words spelled correctly, no account being taken of the difficulty of the word spelled. The score is an index of the total number of the non-technical words of the English language which the pupil can spell correctly.

Starch's spelling lists. Lists I and II are given on the following pages. In using these lists as tests the words are simply dictated, and the pupil allowed to focus his attention upon the spelling. To secure a reliable measure both lists should be used and the two sets of scores averaged.

STARCH'S SPELLING LIST, No. 1.

1. add	5. rat	9. cart
2. but	6. sun	10. come
3. get	7. alum	11. easy
4. low	8. blow	12. fell

¹ Ayres, L. P. *Measurement of Spelling Ability*, p. 38.

- | | | |
|-----------|--------------|----------------------|
| 13. foul | 42. accrue | 71. flourish |
| 14. gold | 43. bottom | 72. luckless |
| 15. head | 44. chapel | 73. national |
| 16. kiss | 45. dragon | 74. pinnacle |
| 17. long | 46. filter | 75. reducent |
| 18. mock | 47. hearse | 76. standing |
| 19. neck | 48. laden | 77. venturer |
| 20. rest | 49. milden | 78. ascension |
| 21. spur | 50. pilfer | 79. dishallow |
| 22. then | 51. rabbit | 80. imposture |
| 23. vile | 52. school | 81. invective |
| 24. afoot | 53. shroud | 82. rebellion |
| 25. black | 54. starch | 83. scrimping |
| 26. brush | 55. vanity | 84. unalloyed |
| 27. close | 56. bizarre | 85. volunteer |
| 28. dodge | 57. compose | 86. cardinally |
| 29. faint | 58. dismiss | 87. connective |
| 30. force | 59. faction | 88. effrontery |
| 31. grape | 60. hemlock | 89. indistinct |
| 32. honor | 61. leopard | 90. nunciature |
| 33. mince | 62. omnibus | 91. sphericity |
| 34. paint | 63. procure | 92. attenuation |
| 35. prism | 64. rinsing | 93. fulminating |
| 36. rogue | 65. splashy | 94. lamentation |
| 37. shape | 66. torpedo | 95. secretarial |
| 38. steal | 67. worship | 96. apparitional |
| 39. swain | 68. bescreen | 97. intermissive |
| 40. title | 69. commence | 98. subjectively |
| 41. wheat | 70. estimate | 99. inspirational |
| | | 100. ineffectuality. |

STARCH'S SPELLING LIST, No. 2

- | | | |
|---------|----------|-----------|
| 1. air | 9. cast | 17. look |
| 2. cat | 10. corn | 18. mold |
| 3. hop | 11. envy | 19. part |
| 4. man | 12. feud | 20. ruin |
| 5. row | 13. game | 21. take |
| 6. tap | 14. grow | 22. tree |
| 7. awry | 15. home | 23. well |
| 8. blue | 16. knee | 24. allay |

25. blaze	50. portal	75. reformer
26. buggy	51. recipe	76. thorough
27. clown	52. scrape	77. watering
28. doubt	53. simple	78. belonging
29. false	54. strain	79. displayed
30. forth	55. weaken	80. indentation
31. grass	56. breaker	81. mercenary
32. house	57. congeal	82. redevelop
33. money	58. disturb	83. senescent
34. paper	59. foreign	84. uncharged
35. quill	60. hoggery	85. whichever
36. rough	61. meaning	86. centennial
37. shout	62. onerate	87. constitute
38. stick	63. provoke	88. exaltation
39. swear	64. salient	89. invocative
40. trump	65. station	90. personable
41. whirl	66. trample	91. strawberry
42. action	67. abstract	92. concentrate
43. bridle	68. bulletin	93. imaginative
44. charge	69. covenant	94. mathematics
45. driver	70. eugenics	95. selfishness
46. finger	71. friskful	96. collectivity
47. heaven	72. luminous	97. marriageable
48. legend	73. opulence	98. agriculturist
49. motley	74. planchet	99. quarantinable
		100. relinquishment

III. STANDARDS

The Ayres Scale. In classifying the words of his list according to difficulty, Ayres determined the average per cent of the pupils of each grade who spelled the words correctly. Thus the words of column O were spelled correctly by 50 per cent of the third-grade pupils, 73 per cent of the fourth-grade pupils, 84 per cent of the fifth-grade pupils, 92 per cent of the sixth-grade pupils, 96 per cent of the seventh-grade pupils, and 99 per cent of the eighth-grade pupils. These per cents, which are printed at the head of each

column, represent the average spelling ability of pupils in the several grades when the words are dictated in lists. When the words are used in timed sentences the averages have been 5 to 15 per cent lower.

In Boston minimum word lists for each grade have been carefully built up on the basis of the words which the pupils use in their written work. When the Boston pupils were tested on the words of the Ayres Scale which are included in their several grade lists, the per cent of correct spellings was conspicuously above the standards given by Ayres.¹ In reporting that study Ballou suggests that "this may be due to the fact that the Boston pupils had been taught the words; whereas, the pupils in the eighty-four cities where Dr. Ayres gave his lists, and on the results of which he standardized the words for his spelling scale, may not have been taught them."

For this reason it may be seriously questioned whether the averages which Ayres gives are satisfactory standards of spelling ability for the foundation words of the language. Ayres says: "Probably the scale will have served its greatest usefulness in any locality when the school children have mastered these one thousand words so thoroughly that the scale has become quite useless as a measuring instrument." In the past we have not had the advantage of such a list and have distributed our efforts in teaching spelling over a very much larger list of words. If we accept these one thousand words as the foundation words of our language, we should place prime emphasis upon teaching them. This being the

¹ Ballou, F. W. "Measuring Boston's Spelling Ability by the Ayres Spelling Scale"; in *School and Society*, vol. v, pp. 267-70.

case a satisfactory eighth-grade standard would approximate one hundred per cent for all of the words. For the preceding grades, the standard would be one hundred per cent for the words of the list which the pupils had been taught. For example, the easiest nine hundred words might be used for the seventh grade, the easiest seven hundred and fifty for the sixth grade, and so forth. The use of the scale in the way Ayres suggests would seem to lead to standards of this type. The distribution of the words among the several grades and the optimum standards must be determined by experimentation.

The Starch tests. Starch gives the following standards for his tests based on their use with over twenty-five hundred pupils.

	Grade							
	I	II	III	IV	V	VI	VII	VIII
Per cent of words spelled correctly....	10	30	40	51	61	71	78	85

These standards are interpreted thus: the average eighth-grade pupil should be able to spell correctly eighty-five per cent of the non-technical words of the English language, or eighty-five of the one hundred words in any one of Starch's tests.

IV. HOW TO LOCATE SPELLING DIFFICULTIES

Locating bad spellers. On page 119 it was stated that a list of ten to twenty words would give a reliable measure of a school system, or a good-sized class, but that a test of

one hundred words or more was necessary to obtain a reliable measure of the spelling ability of individual pupils. From the standpoint of making her instruction most effective the teacher is not so much concerned with securing a reliable measure as in locating the pupils who are below standard, and who for that reason need instruction. A test of twenty words will locate many of the cases, but a test of fifty or one hundred words will be more effective.

A low class average may be due to one or more of three conditions: —

1. The class as a whole may be unable to spell certain words.
2. Certain pupils may be unable to spell a large number of the words of the test.
3. The errors may be rather uniformly distributed as to both words and pupils.

To determine the extent to which each condition causes the low class average, the teacher should make the following type of tabulation from the test papers.

<i>Words of the test</i>	<i>Pupils</i>											
	1	2	3	4	5	6	7	8	9	10	11	12
catch.....	c	-	-	-	c	-	-	-	-	c	-	c
black.....	c	c	c	c	c	c	c	c	c	c	-	c
warm.....	c	c	c	c	c	c	c	c	c	c	c	c
unless.....	c	c	c	c	c	c	c	c	c	c	-	c
clothing.....	c	-	c	-	-	c	-	-	c	c	-	-
began.....	c	c	c	c	c	c	c	c	c	c	-	c

c indicates the word was correctly spelled.

Although these words are listed by Ayres as being equally difficult for pupils in general, they are not necessarily so for

particular pupils. Obviously in the class here represented "catch" and "clothing" need general emphasis, while only certain pupils need to give attention to "black," "began," and "unless." Pupil 11 has misspelled five out of six words, and hence probably is a "poor speller."

Individuality in spelling difficulties. Simply to know that a pupil is below standard in ability is of little value to the teacher, because spelling ability is specific and not general. In general the ability to spell one word does not imply ability to spell another word, nor does the lack of ability to spell a given word indicate that a pupil cannot spell another word. Hence the teacher should make a very careful diagnosis of the spelling ability of each pupil whose test score is below standard to ascertain just what words he cannot spell of those he is expected to spell.

This is accomplished by giving the pupils below standard a test including all of the words which they are expected to be able to spell. Such a test is not for the purpose of measurement but should be thought of as the first step in the teaching of spelling. Each pupil should be required to make from this test a list of all the words which he has spelled incorrectly. The words of this list are the ones he needs to study. It is obvious that to ask a pupil to study words which he can already spell correctly is to ask him to use his time without profit.

"Spelling Demons." Certain frequently used words are very frequently misspelled. Jones¹ has given us a list of 100 words which he found misspelled most frequently in children's compositions. He calls them the "One hundred Spell-

¹ See page 114 for a description of Jones's study.

ing Demons of the English Language." Nine tenths of these words are found in Jones's list for the second and third grade. Four fifths of these words are found in Ayres's list. A teacher will make no mistake in emphasizing these words in the teaching of spelling until the pupils can spell them correctly.

**THE ONE HUNDRED SPELLING DEMONS OF THE ENGLISH LANGUAGE
(Jones)**

which	can't	guess	they
their	sure	says	half
there	loose	having	break
separate	lose	just	huy
don't	Wednesday	doctor	again
meant	country	whether	very
business	February	believe	none
many	know	knew	week
friend	could	laid	often
some	seems	tear	whole
heen	Tuesday	choose	won't
since	wear	tired	cough
used	answer	grammar	piece
always	two	minute	raise
where	too	any	ache
women	ready	much	read
done	forty	beginning	said
hear	hour	blue	hoarse
here	trouble	though	shoes
write	among	coming	to-night
writing	busy	early	wrote
heard	built	instead	enough
does	color	easy	truly
once	making	through	sugar
would	dear	every	straight

Types of misspellings. A pupil's spelling difficulty is not completely diagnosed when the words he does not spell correctly are located. Errors in spelling are seldom if ever distributed uniformly among the several letters composing the word. Neither does it appear that there is much uniformity in the location of errors in different words. Certain

words are misspelled in only a few ways, while other words are misspelled in many ways. Certain misspellings occur frequently, while others seldom occur. In Table XVIII the misspelling of certain words found in the papers of 80 seventh-grade pupils are given, together with the frequency of each. The words were taken from column S of the Ayres Scale. Where no number follows the word that type of misspelling of the word occurred but once.¹

TABLE XVIII. THE MISSPELLINGS OF EIGHTY SEVENTH-GRADE PUPILS ON A COLUMN SPELLING TEST

I. affair	sertain	marage
affere	certin	marriege
affire	sectrain	X. mention
afair (2)	IV. difference	mension (8)
affaired	difference (10)	menioned
afer	diffierence	meantion (2)
II. assist	V. examination	menchion
assit (3)	examation (10)	XI. motion
asist (2)	examition	moshen
ascist	examnition	moticem
assest	excamation	motation
assaist	excanitions	montion
assest	examanaition (3)	XII. neither
assiaist	VI. government	neather (6)
acist	goverment (9)	nether
aciist (2)	govament	niether (2)
accisted	governement	nieghter
assantant	gorvement	XIII. opinion
assised	VII. improvement	oppinion (5)
accessise	improvement (7)	opinon (2)
accest	impovement	opinton
astist	VIII. investigate	oppoinen
assis	investagate (3)	oppinum
assite	envesigatige	oppenion (2)
III. certain	investiage	opion (3)
certian (7)	IX. marriage	oponion (2)
serten	marrage (5)	oppion (2)

¹ See also Sears, J. B., *Spelling Efficiency in the Oakland Schools*. Board of Education Bulletin, Oakland, California, p. 51.

XIV.	particular		perciular	possibbe
	particuliar		pectuliar	posible
	particular		pecticular	XVI. serious
	partictuler		pertictultural	cyreaus
	perticular (8)	XV. possible	patuclure	cerrious
	particlar		peculiar	scerious
	pertucular		pectulair	serrious (2)
	partcular		possible (4)	cerious
	partiular		possible	sereaus
	partular		posable	XVII. stopped
	paticular		posiable (2)	stoped (13)
	partular		possiable (5)	stopts
	paticular		posobile	stocted

Teaching the pupil to correct his errors in spelling. Spelling consists in forming correct and fixed associations "between the successive letters of a word and between the word thus spelled and the meaning."¹ The laws governing the formation of fixed associations are those of habit formation. The first step in habit formation is to get the attention of the child focused upon the associations to be formed. The second step is to secure sufficient repetition. Repetition of the associations is secured both through drill and through using the word in written expression. The pupil must give attention to the repetitions of the associations in order to insure that wrong associations will not be made.

Numerous experiments have shown that pupils can spell correctly a large per cent of the words in the lists in spellers before they have studied them. Because of this fact the assignment of the spelling lesson should include the dictation of the words to the pupils so that each might know what words he needed to study. The teacher would also learn what words she should emphasize in her instruction.

¹ Freeman, F. N. *Psychology of the Common Branches*, p. 115.

Some writers state that a pupil should not be permitted to spell a word incorrectly when it can be avoided, and for this reason pupils should learn to spell words correctly before they are required to write them. Just how important it is to do this we do not know. In certain cases it appears that a child or an adult learns to spell certain words correctly by having his attention directed to his errors. The fact of his error serves to direct his attention to learning to spell the word correctly. Those who believe that evil effects will come from having pupils write words which they cannot spell correctly may direct them to omit those words which they think they cannot spell correctly.

The dictation of the words in assigning the spelling lesson, together with the detailed testing of the pupils as suggested on page 131, reveals to the teacher the words upon which she must exercise her ability as a teacher of spelling. It also reveals to her the pupils to whom instruction should be directed in the case of each word. Particular methods and devices by means of which the laws of habit formation may be fulfilled are described in books which deal with the teaching of spelling.¹

Causes of some misspellings. Certain associations in the spelling of a word appear to be more crucial than others. Take, for example, the word "examination," the fifth word in Table XVIII. The letters *e-x* and *t-i-o-n* were correctly associated in every instance. All of the errors occurred in the other syllables. A study of Table XVIII shows that certain forms of misspelling occur more frequently than others,

¹ A very good chapter (vi) will be found in Freeman's *Psychology of the Common Branches*. See also Cook and O'Shea, *The Child and his Spelling*.

and that most of the misspellings may be attributed to certain specific causes. Forms of misspelling such as "particular," "partuler," "opinon," "impovement," "possibbe," are probably due to carelessness or accident. Relatively few of the misspellings in Table XVIII may be assigned to this cause. Errors of this type probably cannot be entirely eliminated from uncorrected manuscript. However, drill will reduce the number of such errors to a satisfactory minimum.

A more prolific source of error is mispronunciation of the word by the pupil. He may have acquired this from the teacher, but more likely from those with whom he associates outside of school. Or it may have been acquired from lack of attention to the form of the word. Such misspellings as the following are probably caused by mispronunciation: "perticular," "particlar," "investagate," "goverment," "examation."

A very striking instance of this type of spelling error and its cause came to the attention of the writer a few years ago. A man who had taught geometry for a number of years used the word "frustum" in a manuscript, spelling it "frus-trum" which agreed with his pronunciation of the word. This manuscript was read by a number of well-known mathematicians who read it critically. Only two noted the misspelling of the word, and one mathematician, who took much pride in his ability to spell correctly and who was the author of several textbooks, admitted that he had always pronounced and spelled the word "frustrum."

Other errors listed in Table XVIII are due to certain phonic irregularities of the English language, for example, certain misspellings of "assist," "certain," "affair," "marriage,"

"motion," "neither," and "serious." Still other errors, such as, "stoped," and "improvment," are due to certain silent letters. In a few cases it appears that the pupil was not acquainted with the form of the word. Such cases probably should not be counted as errors but rather as words unknown.

Good teaching of spelling. In teaching the spelling of a word the child's attention should be directed to the crucial associations. If the word is one like "government," his attention should be called to the correct pronunciation. If it is such a word as "their," his attention should be called to the use of the word. To eliminate spelling errors a pupil's attention should be called to his particular error and he should be helped to remove the cause. If the cause is mispronunciation, see that he learns to pronounce the word correctly. If the error is due to a confusion of letters the pupil should be given some device to prevent this confusion. The following is a device which may be used for especially difficult words: —

Par-tic-u-lar

I frequently misspell _____ in writing compositions but now I am going to learn to spell it correctly. My teacher tells me that I do not look at the _____ syllables and letters closely enough. I am going to do it now with _____ care. I see that the word has _____ syllables. The first syllable is _____. The vowel of this syllable is _____, the first letter of the alphabet. The last syllable is _____ and the vowel is also _____. The word contains _____ letters, the other vowels are _____ and _____. Now that I have looked at the word carefully I am going to be very _____ in spelling it. I am also going to be _____ in pronouncing it. I am going to remember that the vowel in the first syllable and in the last syllable is an _____. I am not going to pronounce those syllables as if the vowel were *e* instead of _____. I am going to be

very —— about both spelling and pronouncing this word. I want it to be correct in every ——.

This device is used by providing the pupil who needs instruction with a printed or typewritten copy. The pupil is required to fill in the blank spaces correctly. This is repeated until the correct associations are fixed.

Devices for improving spelling. The following device serves to direct the pupil to see his errors in a wholesome way. It has yielded very gratifying results in the Training School of the Kansas State Normal:¹

When the spelling sentences or lists have been written each pupil is required (1) to mark each word, the spelling of which he doubts; (2) as far as possible he is encouraged to test the validity of his doubts by known means outside of the dictionary, finally checking up all doubted words by using the dictionary; and (3) he then writes all of the misspelled words, which he has thus detected, correctly spelled in separate lists; (4) at this point the pupils' papers are exchanged, the teacher spelling all words and the pupils marking those found to be misspelled on the papers; and finally (5) when the papers are returned to their owners the additional misspelled words discovered should be added to their individual lists.

The pupil's spelling is scored by the teacher on the basis of the correctness of his doubts as well as upon the number of words spelled correctly. In the absence of a scientific determination of the relative significance of *spelling of words correctly* and *doubting correctly* the same value is assigned to

¹ Lull, Herbert G., "A Plan for Developing a Spelling Consciousness"; in *Elementary School Journal*, vol. 17, p. 355.

each. The pupils are scored both for doubting words spelled correctly, and for not doubting words spelled incorrectly.

Making associations automatic. Getting the pupil to spell a word correctly is only the first step. There must be attentive repetitions of the correct associations until they have become automatic. In this respect spelling is similar to arithmetic. In the teaching of the operations of arithmetic drill occupies a prominent place, but in the case of spelling our teaching has been confined primarily to testing pupils. Requiring pupils to write each misspelled word ten or twenty times is an effort to provide practice. Such practice is unsatisfactory. After the first writing of the word the pupil probably copies. Hence the repetitions are not attentive.

Practice upon words which are misspelled by a majority of the pupils can be secured by having them recur in the spelling lesson from day to day. This plan provides the same drill for all pupils regardless of whether they misspell the word or not. In this respect it is unsatisfactory.

Courtis's spelling practice tests. In order to provide each pupil with the practice he needs Courtis has devised a series of practice tests in spelling similar to those for arithmetic. A lesson of the practice tests consists of a story with the words to be spelled printed in heavy-faced type. The pupil is directed to study these words. On the reverse side the story is printed with the spelling words omitted. A specimen lesson follows.

DETROIT PUBLIC SCHOOLS

PRACTICE TESTS IN SPELLING

Lesson No. 4. — A Trip to a Great City

On the Sixth of December we left the Green Mountains of Vermont to visit an uncle, who had just returned from a long voyage.

He had acquired immense wealth by deals in leather and turpentine. We were glad to arrive in Chicago on the eleventh for we had often dreamed of this visit. But how deceiving are dreams. Instead of being met at the depot as we expected, we found we did not know a single man or woman in all the great crowd in the station that evening.

We had to argue with ourselves to try to understand how anyone could disappoint us on such an important occasion.

Instructions: Read the paragraphs above and study the spelling of the words printed in heavy type until you can fill in all the blanks on the other side of this sheet correctly in four minutes.

DETROIT PUBLIC SCHOOLS

PRACTICE TESTS IN SPELLING

Lesson No. 4 — A Trip to a Great City

On the	December	we left the	1
Green	of		2
..... to visit an uncle, who had just re-	turned from a long.....		3
He had acquired immense wealth by deals in and		4
..... are dreams. Instead of being met at	the depot as we expected, we found we did not		5
know a single man or woman in all the great	crowd in the station that.....		6
.....		7
We were glad to arrive in.....	for we had often		8
on the	dreamed of this visit. But how.....		9
.....		10
We had to	with ourselves		11
to try to how anyone		12
could us on such an		13
..... occasion.		14

Scores Number Tried Number Right
Name Grade Room

Besides providing each pupil with the practice which he needs, and thus for individual progress, the tests have the added advantage of having each word appear in an appropriate context. A definite time is allowed, and this has been chosen so that a pupil must be able to spell the words automatically when he does the test satisfactorily.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. In what respects does the problem of measuring ability in spelling differ (1) from the problem of measuring ability in arithmetic and (2) from the problem of measuring ability in reading?
2. Compare the method Ayres used in selecting words for his scale with the method used by Buckingham in selecting words for his scale. Which method is the better? Why?
3. Measure the spelling ability of the pupils of your class by means of a timed sentence test and then dictate the test words as separate words. Compare the two sets of scores.
4. Teachers frequently tell with pride that all but two or three of their pupils make a "grade of 100" on a certain test. Should the fact be a cause for a feeling of satisfaction? Were the pupils really tested?
5. Dictate the words for the next spelling lesson before the pupils have studied them. Have each pupil make a list of the words which he misspells and also of the particular misspellings which he has used. Direct the pupils to base their study upon these lists.
6. How could you determine whether the method suggested in question 5 is a good one?
7. Construct a series of "timed sentence spelling tests" for the elementary school, using suitable words from the Ayres Scale.
8. How do the scores obtained by using Starch's Tests differ in meaning from the scores obtained by using a test made from the Ayres Scale?
9. Why does a test of easy words fail to give a measure of spelling ability?
10. Why must the relative difficulty of the words of a test be known if accurate measures are desired?
11. Make a study of the ways in which your pupils misspell. Also ascertain the causes for these misspellings.
12. How can you use this information in making your teaching of spelling more effective?

BIBLIOGRAPHY

Only the most important references are given here. Additional references will be found in the footnotes of the chapter.

1. *Ayres's Spelling Scale*. Copies may be purchased from the Division of Education, Russell Sage Foundation, New York City.

REFERENCES: Ayres, L. P. *A Measuring Scale for Ability in Spelling.* (Division of Education, Russell Sage Foundation, New York City.)

Sears, J. B. *Spelling Efficiency in the Oakland Schools.* Report of the Oakland Spelling Investigation of October, 1914. Oakland, California, Board of Education, 1915. (Bureau of Information, Statistics, and Educational Research, Oakland School Department, Publication no. 1, 1916.)

2. *Buckingham's Spelling Scale.* The scale is not published separately, but may be found in the monograph, — *Spelling Ability; Its Measurement and Distribution*, by B. R. Buckingham. This may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City. (Teachers College Contributions to Education, no. 59.)

REFERENCES: Lewis, E. E. "Testing the Spelling Abilities of Iowa School Children by the Buckingham Spelling Tests"; in *Elementary School Journal*, vol. 16, pp. 556-64. (June, 1916.)

Thorndike, Edward L. "Means of Measuring School Achievements in Spelling"; in *Educational Administration and Supervision*, vol. 1, pp. 306-12. (May, 1915.)

Tidyman, W. F. "A Descriptive and Critical Study of Buckingham's Investigation of Spelling Efficiency"; in *Educational Administration and Supervision*, vol. 2, pp. 290-304. (May, 1916.)

Buckingham's Spelling Tests were used in the Survey of the Gary and Prevocational Schools of New York City. See *Seventeenth Annual Report of the City Superintendent of Schools*, New York City, 1914-15.

3. *Courtis's Standard Research Tests in Spelling.* Copies may be obtained from S. A. Courtis, 82 Eliot Street, Detroit, Mich.
4. *Starch's Spelling Tests.* For these lists and the directions for giving them consult *Educational Measurements*, by Daniel Starch, chap. vi. (The Macmillan Company.) Copies of the tests may be obtained from Daniel Starch, Madison, Wisconsin.

REFERENCE: Starch, Daniel. "The Measurement of Efficiency in Spelling, and the Overlapping of Grades in Combined Measurements of Reading, Writing, and Spelling"; in *Journal of Educational Psychology*, vol. 6. (March, 1915.)

GENERAL REFERENCES

Ayres, Leonard P. *The Spelling Vocabularies of Personal and Business Letters.* (Russell Sage Foundation, Bulletin.)

Cook and O'Shea, *The Child and His Spelling.*

Houser, David J. "The Relation of Spelling Ability to General Intelligence and to Meaning Vocabulary"; in *Elementary School Journal*, vol. 16, pp. 190-99. (December, 1915.)

Jones, N. Franklin. *Concrete Examination of the Material of English Spelling.* (University of South Dakota, Bulletin, 1913.)

144 EDUCATIONAL TESTS AND MEASUREMENTS

Otis, Arthur S. "The Reliability of Spelling Scales, involving a 'Deviation Formula' for Correlation"; in *School and Society*, vol. 4, pp. 676-83, 716-22, 650-56, 793-96. (October 28, November 4, 11, 18, 1916.)

Rice, J. M. "The Futility of the Spelling Grind"; in *The Forum*, vol. 23, pp. 163-72, 409-19.

Tidyman, W. F. "A Critical Study of Rice's Investigation of Spelling Efficiency"; in *Pedagogical Seminary*, vol. 22, pp. 391-400. (September, 1915.)

Wallin, J. E. W. *Spelling Efficiency in Relation to Age, Grade, and Sex, and the Question of Transfer*. (Warwick & York, Baltimore, Md.)

CHAPTER V

HANDWRITING

I. THE PROBLEM OF MEASUREMENT IN HANDWRITING

HANDWRITING is measured in several ways. Frequently the teacher watches the pupil write and passes judgment on one or more of such factors as position, movement, and apparent ease of production. Many teachers examine the script or specimen of handwriting, seeking to discover the merit of the factors which enter into its production. Other examinations of the specimen seek to estimate the general merit or quality of the handwriting, ignoring those factors which enter into its production. These estimates are ordinarily made with no specific factors in mind other than rather vague ideas of good appearance, beauty, legibility, a good business hand, and the like. These methods result in inaccurate and unsatisfactory measures of handwriting.

Another method of measuring. The problem of measurement in handwriting differs fundamentally from that of any of the school subjects treated in the preceding chapters. The answer to an example or the spelling of a word is either right or wrong. Typical specimens of handwriting cannot be definitely classified as either legible or illegible. Instead there are degrees of legibility and these cannot be easily defined. A specimen of handwriting which is read with great difficulty, if at all, by one person, is read with considerable ease by another. Thus, legibility depends in part upon the reader as well as upon the form of the handwriting.

Many features of form, such as the size of the letters, the form of the letters, the spacing of words, the kind of pencil or pen used, and the like, affect legibility. Legibility concerns the reader. From the standpoint of the writer the speed and ease of production are the significant features. The problem of measurement will be treated in more detail under the topics of speed and of quality.

II. HANDWRITING SCALES

Measuring speed. Speed is measured by requiring the pupil to write for a specified time under standard conditions, and then counting the letters written. The speed is stated in terms of the number of letters written in a minute, or the average time of writing one letter. When measuring speed a two or three-minute period should be allowed, taking the time by means of a stop-watch or the second-hand of an ordinary watch. Obviously the teacher should see that all pupils are well provided with good pen-points, ink, and paper, unless they use pencils, in which case there should be a sufficient supply of well-sharpened pencils. The directions to the pupils should be given with due precautions against misunderstanding.

Pupils should be asked to write a suitable selection which they have memorized. To guard against lapses of memory, the pupils should be asked to repeat in concert the selection to be used. If convenient it is well to provide each pupil with a printed or typewritten copy of the selection. When this cannot be done the selection may be written on the blackboard where all can see it. The selection should contain no words which the pupils cannot spell readily. It is well to

have them practice writing the more difficult words before the test is begun. Do not use material which the pupils must compose as they write, for this would be worthless in testing. The rate of writing unfamiliar material from a printed copy will vary with the pupil's rate of reading and so will not give a true measure of speed. Dictated material should be used only when the teacher wishes to control the speed, not when speed is to be measured.

Selections for the speed tests. Different investigators have required pupils to write different material. Several have used the first line or the first stanza of the poem "Mary had a little lamb." "Sing a song of sixpence" has been used. Other sentences which have furnished copy are: "Jolly kings bring gifts while happy maids dance." "A quick brown fox jumps over the lazy dog."¹ "Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card. John vanished behind the bushes and the carriage moved along down the driveway."² In the Cleveland Survey the first three sentences of Lincoln's Gettysburg Address were written, and Ayres has used this selection in the "Gettysburg Edition" of his scale. In several surveys the pupils were allowed to write any familiar stanza of a poem. The chief principles to bear in mind in selecting materials are: first, to use material in the lower grades which will not furnish difficulties in spelling and remembering; and second, to use material which will be uniform in all classes which are to be compared.

¹ This sentence was used in securing specimens for the Freeman Scale. It contains all of the letters of the alphabet.

² These sentences were used in securing the specimens for the Thorndike Scale.

The following directions are representative of many which have been used with good results:—

“ Write the stanza of the poem which you have learned. When you have written the stanza, write it again, and keep on writing until I tell you to stop. Write as well as you can and as fast as you can. Write on one side of the paper. When you fill one page, use another. Place your paper in position and see that your pen and ink are ready. When I say ‘ready,’ ink your pen and place your hand in position to write, but do not begin until I say, ‘Start.’ When I say, ‘Stop,’ all stop at once and raise your hands so I can see that you have stopped. Remember: Fast work and good work. Ready! Start!” At the end of three minutes, “ Stop! ”

Measuring quality; use of scales. The “quality” of handwriting is generally measured by means of a scale, which consists of a number of specimens of handwriting arranged in order of merit or legibility. The scales in most general use are the ones constructed by Thorndike¹ and by Ayres.²

Thorndike constructed his scale on the basis of three characteristics — beauty, legibility, and general merit. The degree of these characteristics represented in the specimens of the scale was determined by the consensus of opinion of competent judges. *Ayres* constructed his scale on the basis of legibility alone. He defined legibility in terms of ease of reading. That specimen was defined as most legible which was read most easily. The numerical values of the specimens of the *Thorndike Scale* range from 4 to 18, one or

¹ *Thorndike, E. L. “Handwriting”; in Teachers College Record, vol. 2, no. 5. (March, 1910.)*

² *Ayres, L. P. A Scale for Measuring the Handwriting of School-Children. (Russell Sage Foundation, Bulletin no. 113.) Ayres has also constructed an Adult Scale and the “Gettysburg Edition.” In this book the term, Ayres’s Scale, refers to the “Three Slant Edition” unless otherwise noted.*

Quality 11.

John vanished behind the bushes and the carriage moved along down the driveway. The audience

driveway. The audience of passers-by, which had been gathering about them melted away in an instant leaving only a poor old lady on the curb. Albert was sadly striking.

Quality 9.

Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card, John vanished behind the bushes which had been gathering about them melted away in an instant leaving only a poor old lady on the curb. Albert was sadly striking.

Then the carelessly dressed gentleman stepped lightly into Warren's carriage moved and held out a small card, John vanished

Quality 8.

Moved along down the driveway. The audience of passers-by which had been gathering about them melted away.

Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small card, John vanished behind the bushes and the carriage moved

FIG. 7. A SECTION OF THE THORNDIKE HANDWRITING SCALE

(Reduced $\frac{1}{2}$ in size.) Quality 9 of this scale is approximately equal to quality 40 of the Ayres Scale. Quality 11 is better than the Ayres 50. See table of relative values, on page 169.

40

The appearance of Ryo,
beard, his rusty fowlir
and the arm of name I
who demanded that ever
wilder ment of his misla.

The hair of the affrighted peda
terror What was to be done?
him within sink to bays
other his quickened however
leaving of hopes in steed hi

The gallant hero now sp
hour at his toilet, brushes
his best country the in,
looked he as down bro
had and rider furious a ce

FIG. 8. TWO SECTIONS OF THE
(Slightly reduced in size.) These two sections

50

Fain would I pause to dwell,
burst upon the enraptured g.
every to justice ample did I
great so in not was here ou
on get to eager too am ai

His school was a low building
constructed of logs the window
partly strong the of those of
the of back off burden the ta
studying pupils his of mu

Ichabod pride himself,
much as upon his vocal
not a fibre about explo
himself make to recollect
and fiction becoming!

more specimens being given for each degree of quality. A section of the Thorndike Scale is shown in Fig. 7.

Ayres's Scale consists of three types of specimens, vertical, semi-slant, and full slant. Each of these three types is represented by eight degrees of quality to which are assigned the numerical values 20, 30, 40, up to 90. In using this scale it must be remembered that these values are not the same as the per cents used in reporting "grades." A section of the Ayres Scale is reproduced in Fig. 8.

Ayres¹ later devised a scale from specimens of handwriting written by adults. Trained judges used the "Three Slant Edition" in selecting the specimens and in determining their values. This "Adult Scale" is similar to the "Three Slant Edition" in its general plan. Very recently (1917) Ayres devised a third scale, the "Gettysburg Edition." This scale differs from the others in the following particulars. It has one specimen for each step. The specimens are written on ruled paper. The copy is the same for all specimens. In addition to the standardized specimens of the Scale, this edition has directions for securing specimens from a class and for scoring these specimens. It also furnishes standards for speed and quality of handwriting for the grades above the fourth. Ayres asserts that the purpose of these changes is "to increase the reliability of measurements of handwriting."

Johnson and Stone² have made a scale similar in general plan to the Ayres and Thorndike Scales, but based on several

¹ Ayres, L. P. *A Scale for Measuring the Handwriting of Adults.* (Russell Sage Foundation. Bulletin E 138.)

² Johnson, George L., and Stone, C. R. "Measuring the Quality of Handwriting"; in *The Elementary School Journal*, February, 1916.

factors, including movement and a detailed analysis of legibility. Each specimen of the scale is accompanied by a legend which states its defects and merits in terms of the analysis appended, which includes seven factors:—letter formation, uniformity of slant, uniformity of alignment, spacing, quality of line, size, and degree of slant.

Breed and Downs,¹ offer a handwriting scale obtained from a survey of the handwriting of the public schools of Highland Park, Michigan. The specimens collected were scored by using the Thorndike Scale. Specimens were then selected for a five-step scale for each of the following grades, 3d A, 3d B, 4th A, 5th A and 6th A. Values are assigned to these steps in terms of the values on the Thorndike Scale. A standard for speed is given for each grade. This scale furnishes an excellent example of what may be done in the constructing of a scale for local use from specimens collected from the schools concerned. Beyond this the scale has nothing which makes it a rival of the other scales for general use.

*Freeman's*² Scale differs from the other scales in an important respect. It is in reality five scales, one for each of the following characteristics of handwriting: uniformity of slant, uniformity of alignment, quality of line, letter formation, and spacing. These five scales are now printed on one sheet of paper or chart, and each scale is called a division.

¹ Breed, F. S., and Downs, E. F. "Measuring and Standardizing the Handwriting in a School System"; in *Elementary School Journal*, vol. 17. (March, 1917.)

² Freeman, F. N. *The Teaching of Handwriting*. (Houghton Mifflin Company, 1915.) Also, "An Analytical Scale for the Judging of Handwriting"; in *The Elementary School Journal*, vol. 15, p. 432. (April, 1915.)

This scale is very useful in the diagnosis of the handwriting of a pupil.

The score card for detailed analysis. The score card represents another attack upon the problem of measurement. Such instruments as the Ayres and the Thorndike Scales do not require that the user make a detailed analysis of general merit or quality. The score card requires that the essential elements of handwriting be selected and each assigned a value. The score card devised by Gray¹ weights the value of each of the essential elements of handwriting so that the highest value which can be assigned to slant is 5, while spacing of letters may receive 18, neatness, 13, etc. (See Fig. 9.) The use of this score card by teachers in their grading of handwriting would undoubtedly tend to direct their attention to the individual needs of the pupils. So far there is no evidence to show that the use of the score card will result in more accurate measures than the use of any one of the scales. Some claim that the elements of handwriting have not been correctly evaluated. However, it has the advantage that its use trains the user in the analysis of handwriting. Gray well defends the device by saying that agriculturists have long used such score cards to secure very satisfactory and accurate results in judging grain and live stock.

The scales classified as to use. The instruments for measuring quality of handwriting may be classified according to their use into two groups. The Thorndike, Ayres, Johnson-Stone, and Breed-Downs Scales are used

¹ Gray, C. Truman. *A Score Card for the Measurement of Handwriting.* (Bulletin of the University of Texas, no. 37, July, 1915.)

Pupil..... Age..... Date.....
 Grade..... School.....
 Sample Number..... Teacher.....

Sample	Perfect score	Score												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1. Heaviness.....	3
2. Slant	5
Uniformity														
Mixed														
3. Size	7
Uniformity														
Too large														
Too small														
4. Alignment.....	8
5. Spacing of lines	9
Uniformity														
Too close														
Too far apart														
6. Spacing of words	11
Uniformity														
Too close														
Too far apart														
7. Spacing of latters.....	18
Uniformity														
Too close														
Too far apart														
8. Neatness.....	13
Blotches														
Carelessness														
9. Formation of letters.....	(26)
General form	8
Smoothness.....	6
Letters not closed.....	5
Parts omitted.....	5
Parts added.....	2
Total Score.....	

FIG. 9. STANDARD SCORE CARD FOR MEASURING HANDWRITING
 (Devised by C. T. Gray)

to measure general quality as a unit. The Freeman Charts and Gray Score Card are used to analyze general quality into its factors. Each factor may be measured and the values placed on the several factors may be added together to furnish a general measure. There is no evidence that such a general measure is any more accurate than one which is more easily secured by means of one of the general scales. Hence it is probable that the first group of scales will remain in use where general measurement or surveys are desired, and that the Freeman Charts and Gray Score Card will be used for what we shall call diagnostic measurement.

Methods of using scales. The quality of a specimen of handwriting is measured by comparing it with the specimens which compose the scale. Its value is the scale value of the specimen of the scale which it most nearly resembles. Several methods of comparing specimens with a scale are in vogue. When a teacher works independently the best method is the sorting method described by Ayres,¹ as follows: —

The scorer sorts into separate piles all of the papers to be rated, putting in one pile those which he judges to be of quality 20, in another, those which he judges to be of quality 30, and so on for all of the different qualities. He then carefully compares all of the papers in each pile with each other and with the samples of that value reproduced on the scale, so as to make sure that he has not included in the pile any samples that might more justly be assigned to the next higher or lower piles.

Another method, which requires less time, but does not secure as good results as the sorting method, is the ascend-

¹ Ayres, L. P. *A Scale for Measuring the Quality of Handwriting of Adults.* (Russell Sage Foundation, Bulletin E 138, p. 9.)

ing-descending method. This requires that the specimen being examined be moved from the lowest step on the scale toward the higher steps until the judge decides the specimen on the scale to be superior to the specimen in hand. Then, beginning at the upper end of the scale, the specimen must be compared with the steps of the scale until the judge decides the specimen on the scale to be inferior to the one in hand. The specimen in hand then receives the rating represented by the point midway between the step of the scale reached in the ascending and the step reached in the descending series of comparisons. For example, working upwards on the Ayres Scale the judge stops at 70 and working downwards at 60. The specimen in hand would then be rated 65.

This method may be varied by rating all of the specimens of a class by working up the scale, recording the judgments on the back of the specimens rated. Next, rate all the specimens working down the scale, and record the judgments on the face of the specimens. Finally, the judge goes over the specimens and establishes the midway point for each specimen as the true rating.

Whenever three or more persons can work together in scoring specimens the results may be expected to be more satisfactory than those secured by independent work. All the members of the group should examine the specimen of writing and confer concerning the rating it should receive. A majority of the group must agree before a score is assigned to the specimen.

A method which will require more time, but one which will secure more accurate results than the methods described

above, is one in which a group of three or more persons score the specimens independently, using the sorting method. Then the scores assigned by all of the judges to a specimen are averaged and the result taken as the true score for that specimen. The accuracy of the resulting scores will increase with the size of the group of judges. All of these methods may be used with either the Ayres or Thorndike Scales, and with modifications with the other scales.

The results of a number of investigations have shown that careful training in a relatively poor method of using a scale will produce a marked improvement in the accuracy of the scores. It should follow that careful training in the use of the sorting and group methods would produce highly accurate results.

Measurement for diagnosis. In using Gray's Score Card and the Freeman Scale, measures of each of the several factors concerned in a pupil's handwriting are secured. A record of successive measurements show the progress or decline in the general quality of the pupil's writing, and thus furnish a basis for class marks. But far more important, it will show just which abilities have not been sufficiently improved. These abilities will then be the points of attack for the teacher and pupil in their subsequent work. For example, a record as shown on the Gray Score Card might indicate that a pupil's handwriting was suffering chiefly because of poor letter formation. A closer inspection would show that letter formation was very often defective in two items, letters not closed and parts omitted. Such diagnosis reveals a definite problem for the teacher.

Use of the score card. The score card (see page 155) may be used for a pupil, or a class. If it is used for a pupil, the numerals along the top may be taken to indicate weeks, months, or other intervals. In the column under the numeral 1 the first scores of a pupil's handwriting should be entered. A month later a second series of scores should be entered in the column headed by the numeral 2. The next month another series of scores should be entered under numeral 3, and so on. At the close of a term there will appear a very useful record of the child's experience in the learning of handwriting. This use of the score card Gray calls a clinical study.

If the card is used for a class, the numerals at the head of the columns stand for the specimens written by the several pupils of the class. The totals at the bottom will furnish an interesting comparison of the ability of the pupils. Each pupil knowing his number can tell how he stands in relation to the other members of the class. If a new score card is posted each month, a pupil may see whether he is gaining or losing in his position in the class. If he is losing, he will be inclined to seek the reason. He may see that his neatness has a low score. This furnishes a strong incentive for work to improve in neatness. Teachers and supervisors might compare their records. The use of the card may be varied by training pupils to score their own or others' handwriting, or by one teacher calling on another teacher to score the handwriting of her pupils.

The individual record card shown in Fig. 10 is a very simple form of a score card designed to be used with the Freeman Scale.

Pupil's Name..... City

	First trial Date.....	Second trial Date.....	Third trial Date.....	Fourth trial Date.....	Teacher.....	Building.....	Grade.....	Age.....
Chart I (Slant)								
Chart II (Alignment)								
Chart III (Quality of line)								
Chart IV (Letter formation)								
Chart V (Spacing)								
Total (value on Freeman Scale)								
Quality (value on Ayres Scale)								
Speed (Letters per minute)								

FIG. 10. INDIVIDUAL RECORD CARD, FREEMAN SCALE

The Freeman Scale.¹ The first of the five divisions of the Freeman Scale represents three degrees of uniformity of slant. In using this division, as in using the next division, judgments will be made more easily if a slant and alignment gauge is used.² The second division represents uniformity of alignment. The user must be careful to note that letters which are close together show deviations in alignment more prominently than letters written farther apart.

¹ Described more fully by Freeman in *Teaching of Handwriting*. (Houghton Mifflin Company, 1914.)

² Freeman, F. N. *The Teaching of Handwriting*, p. 151. The slant gauge consists of three rows of parallel lines. The lines in one row are vertical and in each of the other rows the lines are set at a uniform slant. The alignment gauge consists of one straight line four or five inches long. These lines may be drawn on transparent paper and placed over a specimen of handwriting to assist in determining the deviations from uniformity in slant and alignment.

The third division shows the quality of line or stroke. A reading-glass will aid in judging with this division. The fourth division is intended to measure letter formation. Freeman describes eight illegible forms of letters which should be counted as errors. Two principles should control here: first, whatever slant or type of script the pupil may use, consistency to that choice should be maintained; and second, no letter should vary from its recognized form so much as to be easily mistaken for another letter. The fifth division shows different kinds of spacing. Letters may be crowded or spread too far. The same applies to words.

In each division the three degrees of excellence are given scores of 1, 3, and 5 respectively. The intermediate values of 2 and 4 may also be used. If the old edition of the scale is used, the scores assigned to the specimens of letter formation are 2, 6, and 10. Freeman suggests that the specimens be scored by using the score for letter formation as placed on the new edition of the chart, and then doubling these scores in making up the total score.

Using the Freeman Scale. This scale may be used for measuring specimens from all members of a class, but frequently it is used to measure specimens written by those ranking conspicuously below the average ability or below the standard ability of the class. This needy group of pupils may be selected by the teacher's unaided judgment, but preferably by the use of the Thorndike or Ayres Scales.

Freeman¹ has recently issued the following suggestion for using his scale: —

¹ Freeman, F. N. *Experimental Education*, p. 86. (Houghton Mifflin Company, 1916.)

The specimen to be judged is graded according to each category separately and given the rank of the specimen in the chart with which it most nearly corresponds in each case. The total rank is calculated by summing up the five individual ranks. Thus, if letter formation is given double value, the lowest possible rank is 6 and the highest possible rank is 30 ($5 + 5 + 5 + 10 + 5$), and the range is 24.

Several precautions are to be observed in making the judgments. The value of the method rests upon the fact that different features of the writing are singled out, one at a time, and graded by being given a rank in one of only three steps. The differences between the steps are marked, and the ease of placing a specimen should be correspondingly easy.

This method implies, however, that

- (1) The attention is fixed on only one characteristic at a time.
- (2) The judgment on one point be not allowed to influence the judgment on the other point.
- (3) The same fault be counted only once.
- (4) General impressions be disregarded.

The scores secured by means of the Freeman Scale should be saved to furnish a means of evaluating the results secured from instruction. The scores may be recorded on the specimen, or better on an individual record card, such as shown in Fig. 10. The latter will be more convenient when the teacher wishes to examine a series of scores recorded at intervals over a term of several months.

III. THE RELIABILITY OF MEASURES AND SCALES

Rate and quality contrasted. The measure of a pupil's rate of writing in terms of the number of letters written per minute is definite. It is an accurate measure of the rate at which the pupil wrote when the specimen was secured. It is a true measure of the pupil's normal rate of writing unless the pupil wrote slower or more rapidly than he is accustomed to write. Thus in securing specimens care should be taken

to have the pupils write as nearly at their normal rate as possible.

If we cannot be sure that pupils are writing at their normal speed, we can at least use the standardized instructions for collecting specimens of handwriting. This will insure that the speed at which the pupils write is influenced by the instructions given to them in the same degree as the speed was influenced in establishing the standard. The importance of this safeguard can easily be established if pupils be directed to write a copy as rapidly as they can write it for one minute and after an interval be again asked to write the same copy as well as they can. There will in almost every case be a wide difference in the rates at which they write.

The measurement of quality is different. Some teachers are skeptical of the measures of the quality of handwriting because they believe the scales have not been accurately constructed. It is easy for any teacher to criticize the scales, but it is very difficult for even an expert to improve on them. The construction of a perfect handwriting scale is a task which we must leave to the expert. Meanwhile we shall do well if we make full use of the available instruments. Other teachers are willing to accept the construction of the scales but doubt the accuracy of the scores obtained by using them. These may be answered by the work of Kelly,¹ Lewis,² and Gray,³ who have shown that with equal train-

¹ Kelly, F. J. *Teachers' Marks*, pp. 99-108. (Teachers College Contributions to Education, no. 66, 1914.)

² Lewis, E. E. "The Present Standard of Handwriting in Iowa Normal Training High Schools"; in *School Administration and Supervision*, vol. 1, no. 10, pp. 663-71. (December, 1915.)

³ Gray, C. T. "The Training of Judgment in the Use of the Ayres Scale for Handwriting"; in *Journal of Educational Psychology*, February, 1915, p. 85.

ing in the usual per cent method of grading and in measuring with the aid of a scale, the results of the latter method are more accurate. Lewis asserts that the use of a scale reduces the variation almost one half over the ordinary per cent method. Kelly makes this significant statement: "Teachers have reduced the variability shown in the per cent method by practice at the expense of the children, while they have at the same time decreased their capacity for effective use of the standard scale."

Accuracy of the scores. The accuracy of scores may be considered in either or both of two ways. If one person scores a set of specimens and then, after an interval, scores them again, his scores may be judged to be accurate or inaccurate as they agree or disagree. This lack of agreement is called individual agreement or individual variation. Again several persons may score the same set of specimens, working independently. When their several scores for the same specimen are compared the amount of their agreement is called the group agreement or group variation.

Several investigators¹ have attacked the question of the accuracy of the scores secured by the use of the scales and have found that there is a wide variation in the scores

¹ Breed, F. S., and Culp, Vernon. "An Application and Critique of the Ayres Handwriting Scale"; in *School and Society*, vol. 2, pp. 36-47. (October, 1915.)

Manuel, H. T. "The Use of an Objective Scale for Grading Handwriting"; in *The Elementary School Journal*, vol. 15, no. 5, p. 269. (January, 1915.)

King, Irving, and Johnson, Harry. "The Writing Abilities of the Elementary and Grammar School Pupils of a City School System Measured by the Ayres Scale"; in *Journal of Educational Psychology*, vol. 3, pp. 514-20.

Harvey, Nathan A. "The Use of Handwriting Scales"; in *The American Schoolmaster*, October, 1916, p. 361.

when several judges score the same specimens. These results appear discouraging, but it is significant that these investigators do not condemn the use of the scales as their results would seem to warrant. On the contrary, there appears a faith that scales will be used successfully in the future. This is a confession that the results secured have a limited application.

Training in using the scales. The teacher's ability to make comparisons depends on technique and training. The technique for using scales has been described under the caption of "Methods of Using the Scales." It is significant that none of the investigators referred to used what seems to be the best technique in rating specimens of handwriting. Neither did they use well-trained judges. The two must go together. A teacher might make inaccurate comparisons while using a good method, if poorly trained in that method. The most accurate results will follow from the teacher using the best method after being well trained in the use of that method. Furthermore, there are several methods of training.

Thorndike¹ has proposed a method of training judges in the use of the handwriting scale. He furnishes fifty specimens of handwriting whose value has been determined. The teacher scores each specimen, without referring to its true value. She then compares her score with the true value and notes her error. This is done with each of the fifty specimens. After scoring the fifty specimens once, they are to be taken up in random order and all scored again as before.

¹ Thorndike, E. L. "Teacher's Estimates of Specimens of Handwriting"; in *Teachers College Record*, November, 1914.

Each time they are scored there will be some gain in the accuracy of the scores. Hurt¹ found that three weeks of this training enabled seventeen of twenty-one judges to bring their individual variation within usable limits.

Hurt carefully tested the effects of several methods of training in the use of the Thorndike scale. He has shown that individual independent practice in the use of the scale reduces the individual variation of several ratings of one set of specimens. It was also shown that after two groups of judges had had several weeks of practice, the group which practiced two weeks longer succeeded in still further reducing their individual variation or variations with their own previous ratings. It seems probable that an individual can make his ratings more consistent by a long period of independent practice. But such consistency may not make one judge's ratings agree any better with the ratings of other judges if other judges are not equally well trained.

Hurt used another method of training judges: One group of five judges had practice with informal instruction in the use of the scale for one month. The average of a small group of scores was taken as the true measure of a specimen. These judges rated the specimens independently, but conferred between ratings. These conferences and the instruction given were directed towards the reduction of group variation, hence they did not tend to make a judge any more consistent with his own previous ratings. Neither did the training have a lasting effect. Some judges acquired a greater proficiency in the use of the scale than others.

¹ Hurt, A. O. *A Study of the Reliability of the Thorndike Handwriting Scale.* (An unpublished Master Thesis, University of Missouri.)

This method of training appears to be of doubtful value for the teacher.

Gray¹ tested the effect of practice with instruction in the use of the Ayres Scale. He gave three judges careful instruction and practice for twenty weeks. At the end of ten weeks the group variation was a little more than half of what it was the first week. The twentieth week found their variation reduced to one seventh of the first week's variation. Gray says:—

Accuracy in grading writing by a scale may be produced by careful training in the use of the scale. In the past the assumption has been made that ability to grade expertly in a subject came with an expert knowledge of the subject. While the experiment does not disprove this assumption, it indicates clearly that another avenue of approach to such expert ability is through a period of careful training. This implies that grading may be considered a field more or less by itself, and gives a glimpse of a type of work in education whose chief interest is *the accurate use of units of measurement.*²

Relative values of the different scales. In comparing the Thorndike and Ayres Scales for the reliability of scores secured, Pinter³ finds the Thorndike Scale superior, Starch finds their reliability to be about equal, and Freeman asserts that the Ayres Scale is slightly more reliable. Johnson⁴ found the Thorndike Scale to be better for use in the second and third and possibly in the fourth grades, but considered the

¹ Gray, C. T. "The Training of Judgment in the Use of the Ayres Scale of Handwriting"; in *Journal of Educational Psychology*, February, 1915, p. 85.

² Italics ours.

³ Pinter, R. "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *Journal of Educational Psychology*, vol. 5, pp. 525-31. (November, 1914.)

⁴ Johnson, Joseph Henry. "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *North Carolina High School Bulletin*, vol. 7, pp. 170-73. (October, 1916.)

Ayres Scale more reliable for use in the grades above the fourth. Freeman¹ finds that his analytical scale secures results which are more reliable than the Ayres Scale. These studies are not and do not assume to be conclusive. So it seems fair to conclude that about the same variability may be expected in the use of these three scales.

IV. STANDARD SCORES

Freeman's proposed standards. After measuring the handwriting of a class and thus learning how well the pupils write, the teacher very naturally wishes to know how well they should be expected to write. Several sets of standards of attainment are available to answer this demand. These standards are given in terms of median speed of production and median quality. To compare the scores of a class with these standard scores the teacher will need to find the median of that class.

Freeman² proposed the standards given in Table XIX. The medians for speed are expressed in terms of the number of letters written in one minute. The medians for quality are given in terms of the Ayres Scale.

TABLE XIX. STANDARDS PROPOSED BY FREEMAN

	<i>School grades</i>						
	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Quality.....	44	47	50	55	59	64	70
Speed.....	36	48	56	65	72	80	90

¹ Freeman, F. N. *Experimental Education*, p. 92. (Houghton Mifflin Company, 1916.)

² Freeman, F. N. In the *Fourteenth Yearbook of the National Society for the Study of Education*, part I, chap. v.

To make these standards for quality intelligible to those who have scores secured by the use of the Freeman or Thorndike Scales, these standards are translated into terms of the latter scales in Table XX.

TABLE XX. RELATIVE VALUE OF SCORES ON THREE DIFFERENT SCALES¹

	<i>School grades</i>						
	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Ayres.....	44	47	50	55	59	64	70
Freeman.....	17.9	18.4	19	20	20.8	22	23
Thorndike....	9.36	9.75	10.13	10.76	11.34	11.89	12.66

Table XIX would read thus: A second-grade class should have a median score for speed of 36 letters per minute, and a median score for quality of 44, when scored by the Ayres Scale. Table XX reads, for quality alone, thus: A second-grade class should have a median score of 44 when scored by the Ayres Scale, 17.9 when scored by the Freeman Scale, and 9.36 when scored by the Thorndike Scale.

In the "Gettysburg Edition" of his Scale, Ayres has given standards in the form of distributions of scores for the several grades. These distributions show what we may expect

¹ The scores for the Freeman Scale are taken from graphs and tables given in his *The Teaching of Handwriting*. The scores for the Thorndike Scale are taken from *The Measurement of Efficiency in Reading, Writing, Spelling, and English*. (D. Starch, Madison, Wisconsin, 1914.) For other comparisons, see "Comparable Measures of Handwriting," by L. S. Sackett, in *School and Society*, October 21, 1916; Joseph Henry Johnson, "A Comparison of the Ayres and Thorndike Handwriting Scales"; in *North Carolina High School Bulletin*, 7: 170-73. October, 1916. The comparisons given in these tables may not be statistically accurate, but are as accurate as the present status of measurement demands.

to find, but they do not necessarily show what we should find in an efficient school system. Adherence to the doctrine that the average of present practice constitutes a standard will not carry us far in improvement. Table XXI shows the averages which Ayres gives for the "Gettysburg Edition."

TABLE XXI. STANDARD SCORES FOR THE "GETTYSBURG EDITION"

	<i>Grades</i>						
	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>
Quality	38	42	46	50	54	58	62
Speed.....	32	44	56	64	70	76	80

What these standards represent. Standards of attainment are determined by two considerations: first, they must be attainable by pupils under ordinary school conditions, and without the expenditure of an unreasonable amount of time and effort; second, they should be high enough to assure that the pupil will have sufficient skill in writing to meet the demands which will be made upon him. These considerations are emphasized by the facts that only a limited amount of time is available for the teaching of handwriting in the ordinary school, and that after practice has progressed for a time it does not bring as large returns as it did in its initial period.

The first of these considerations has been met by examining the handwriting of thousands of children, gathered from all parts of our country. Freeman used the results of the scoring of about five thousand specimens from each of the seven

grades. These specimens were selected from a larger number of specimens which were collected in fifty-six large cities of the United States. He found that the average of the scores of the upper half of these specimens gave scores for speed and quality which are approximately the standards he proposes. In checking up the second consideration, Freeman investigated the demands which are made upon those who are employed in several large commercial houses. The returns from this investigation, together with the results of the other investigation, indicated that the standards as proposed are but little more than the minimum essentials. Moreover, Freeman estimates on good evidence that these standards can be attained with an expenditure of not over seventy-five minutes a week.

Other evidence as to standards. Table XXII and Table XXIII give the results of a number of widely-scattered

TABLE XXII. MEDIAN SCORES FOUND (SPEED)

	<i>School Grades</i>							<i>Approximate number of specimens scored</i>
	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>	
Cleveland ¹	60	70	76	80	25,387
Iowa schools ²	39.2	49.2	61.9	65.5	72.6	75	76.5	28,000
Starch's Standards ³	31	38	47	57	65	75	83	4,740
Kansas Medians ⁴	32	35	51	61	67	71	73	6,000
Fifty-six cities ⁵	30.6	43.8	51.2	59.1	62.8	67.9	73	34,000
Freeman's Standards	36	48	56	65	72	80	90	

¹ Judd, Charles H. *Measuring the Work of the Public Schools*. (Report, Survey Committee on the Cleveland Foundation, 1916.)

² Ashbaugh, E. J. *Handwriting of Iowa School Children*. (University of Iowa, Extension Division, Bulletin no. 15, March, 1916.)

³ Starch, D. *The Measurement of Efficiency in Reading, Writing, Spelling, and English*. (University of Wisconsin, 1914.)

⁴ DeVoss, J. C. *Second Annual Report of Bureau of Educational Measurements and Standards*. (Kansas State Normal School, Emporia, Kansas.)

⁵ Freeman, F. N. *Fourteenth Yearbook of the National Society for the Study of Education*, part I. (1915.)

TABLE XXIII. MEDIAN SCORES FOUND (QUALITY)

	School Grades							Approximate number of specimens scored	
	II	III	IV	V	VI	VII	VIII		
Cleveland ¹	45	48	50	55	Ayree	25,387
Iowa ²	35.7	39.8	44.5	49.1	52.3	57	61	"	28,000
Starch's Standards ³ ..	27	33	37	43	47	53	57	"	4,740
Kansas Medians ⁴ ..	44	47	50	55	59	64	70	"	6,000
Fifty-six cities ⁵	39.7	42	45.8	50.5	54.5	58.9	62.8	"	34,000
Freeman's Standards (Ayree's Scale)	44	47	50	55	59	64	70	"	
Salt Lake City ⁶	9.2	10.7	11.1	11.3	12.2	12.8	Thorndike	2,600
Butte, Montana ⁷	8.2	8	8.8	8.9	11.6	11.2	12.1	"	1,400
Southington, Conn. ⁸	10	..	"	
Conneraville, Ind. ⁹	10.3	10	10.3	11.7	11.7	11	"	1,200
Freeman's Standards (Thorndike's Scale)	9.36	9.75	10.13	10.76	11.34	11.89	12.66		

¹ Judd, Charles H. *Measuring the Work of the Public Schools*. (Report, Survey Committee of the Cleveland Foundation, 1916.)

² Ashbaugh, E. J. *Handwriting of Iowa School Children*. (University of Iowa, Extension Division, Bulletin no. 15, March, 1916.)

³ Starch, D. *The Measurement of Efficiency in Reading, Writing, Spelling, and English*. (University of Wisconsin, 1914.)

⁴ DeVoss, J. C. *Second Annual Report of Bureau of Educational Measurements and Standards*. (Kansas State Normal School, Emporia, Kansas.)

⁵ Freeman, F. N. *Fourteenth Yearbook of the National Society for the Study of Education*, part I. (1915.) Revised medians are given in the *Sixteenth Yearbook of the National Society for the Study of Education*, part I. (1916.)

⁶ Report of a Survey of the Schools of Salt Lake City, Utah. (1915.)

⁷ Report of a Survey of the Schools of Butte, Mont., chap. IV. (1914.)

⁸ Witham, E. C. "All the Elements of Handwriting Measured"; in *Educational Administration and Supervision*, vnl. I, pp. 313-24. (May, 1915.)

⁹ Wilson, G. M. "The Handwriting of School Children"; in *Elementary School Teacher*, vol. 6, pp. 540-43. (1911.)

investigations, and show the median scores found in these different places. The Freeman standards are inserted in each table for comparison. The figures in the columns at the extreme right show the total number of specimens rated in each investigation. A comparison of the results shown in these tables with the standards proposed by Freeman (Table XIX), shows that the standards are higher

in most of the cases. This, together with other evidence, points toward a possible modification of the standards set by Freeman.

Standards required for work. Ayres¹ and Ashbaugh² have drawn certain conclusions from the requirements in handwriting which are set up by the examiners of the Municipal Civil Service Commission of New York City. Ashbaugh quotes a letter from the Acting Director of the commission as follows: —

I find that the Municipal Civil Service Commission of New York ordinarily uses the standard of 70 per cent as a passing grade in handwriting, but for positions where handwriting is a special requirement the standard is sometimes set at 75 per cent.

Ayres has shown that the ratings of 70 per cent and 75 per cent, as given by the commission, correspond respectively to scores of 40 and 50 on the Ayres Scale. Since this commission recommends many persons who cannot write better than the 40 specimen of the Ayres Scale, and recommends others who write only as well as the 50 specimen, for positions where handwriting is a special requirement, it would follow that an ability to write as well as 50 on the Ayres Scale would be sufficient for all the demands which many pupils will meet.

There is another obvious demand on the pupil's ability to write. This is the demand made by the high schools and colleges. We have but little data on this point, but many come to high schools unable to write rapidly enough for the

¹ Ayres, L. P. *A Scale for Measuring the Quality of Handwriting of Adults.* (Russell Sage Foundation, Bulletin E 138.)

² Ashbaugh, Ernest J. *Handwriting of Iowa School Children.* (Bulletin of the University of Iowa, March 1, 1916.)

demands placed upon them. They then often sacrifice the quality of their handwriting for the sake of greater speed. Lewis¹ examined the handwriting of 1760 third- and fourth-year students of 166 Iowa Normal Training High Schools. He found their median score for quality to be 59.1 on the Ayres Scale, with a range from 34 to 89. Fifty per cent of the scores fell between 53.6 and 64.3. The average speed of their handwriting was 90 letters per minute. Thus they rank with the seventh-grade standard for quality, and the eighth-grade standard for speed. Comparing their scores with those of many eighth-grade children, as shown in Tables XXII and XXIII, these high-school pupils write from ten to fifteen letters per minute faster, but no better than the average eighth-grade pupil. These data bear out the statement that the higher schools require greater speed of handwriting than the training of the elementary schools has furnished.

V. THE TEACHING SITUATION REVEALED

The point of view which arises when we measure handwriting and set up standards such as those proposed brings out the significance of certain teaching situations. First, it is apparent that handwriting is a very complex ability; second, the teacher of handwriting must see her problem in terms of individuals and not in terms of classes; and, third, these individuals will differ widely in their abilities, in their consequent needs, and in other respects. A brief consideration of these points will make their significance apparent.

¹ Lewis, E. E. "The Present Standard of Handwriting in Iowa Normal Training High Schools"; in *Educational Administration and Supervision*, vol. 1, pp. 663-71. (December, 1915.)

Handwriting a complex ability. The complexity of handwriting ability is seen from two points of view. If we watch the pupil write we see that there are several factors called position, movement, ease of production, etc., and that each of these is reducible to other more elemental factors which are specific muscular adjustments with their attendant conscious processes. When we examine a specimen of handwriting a similar complexity is apparent. Legibility, quality, and speed of production are not simple, but complex. Each in turn may be analyzed into factors which are finally dependent on muscular adjustments and conscious processes, similar to or identical with those reached by analysis from the former point of view. The muscular adjustment, with its attendant conscious process which results from either analysis, constitutes a specific habit or specific ability which must be the point of attack for the teacher.

An individual rather than a class problem. The children which the teacher has before her are not grouped on the basis of their needs in handwriting, but for other administrative and economic reasons. When the needs of a class are analyzed it is found that each individual has an unique equipment of special abilities. Two individuals may receive the same mark or score on their specimen of handwriting, but one specimen may show the need of training along one line of abilities and the other show the need of training of another set of abilities. Hence a general prescription of drills for a class must be very wasteful, since some members will inevitably be fixing habits which are undesirable, and failing to develop the abilities which they need. The teacher has the problem of prescribing those exercises and

engendering those ideas of form which are needed by each individual.

Children differ widely in abilities and needs. Not only will children differ in their individual equipments of specific abilities, but they will differ in their manner of responding to instruction and in their rate of developing through practice. This makes it necessary for a teacher to have not one but several remedies for each shortcoming which is encountered. It also makes it inevitable that after a term of practice and instruction, pupils will still be found to vary widely in their ability to write. Since this is inevitable and due to a characteristic of human nature, the teacher should not seek to secure uniformity of abilities. But the teacher should be concerned that the majority of the class shall have passed a certain milestone in their development, and that every individual not suffering under an unavoidable handicap should make some advance in his ability to write.

Plotting scores, and reading their meaning. The scores secured by measuring the handwriting of a class can be made to reveal valuable information for the teacher. Figs. 11, 12, and 13 show typical conditions. These figures show the distributions of scores in speed and accuracy for three classes. Fig. 11 represents the distribution of scores for a third-grade class. The numerals along the bottom of the figure denote quality on the Ayres Scale, and speed in terms of letters written in one minute. The numerals along the side indicate the number of pupils. A perpendicular solid line shows the location of the median for the class, and a perpendicular broken line shows the location of the standard for that grade. This same general explanation will fit Figs. 12 and 13.

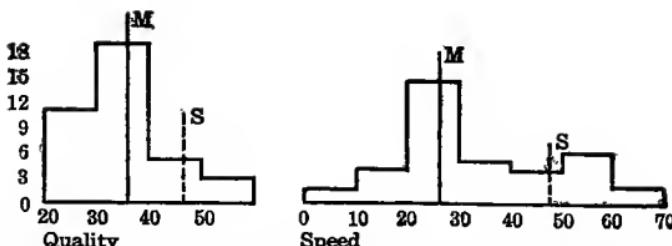


FIG. 11. SHOWING THE DISTRIBUTION OF SCORES IN HANDWRITING OF A THIRD-GRADE CLASS

The line M indicates the median score for the class, the line S the standard for the class.

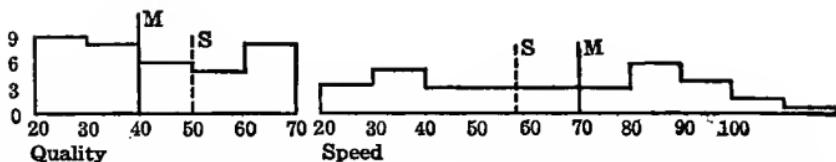


FIG. 12. SHOWING THE DISTRIBUTION OF SCORES IN HANDWRITING OF A FOURTH-GRADE CLASS

The line M indicates the median score for the class, the line S the standard for the class.

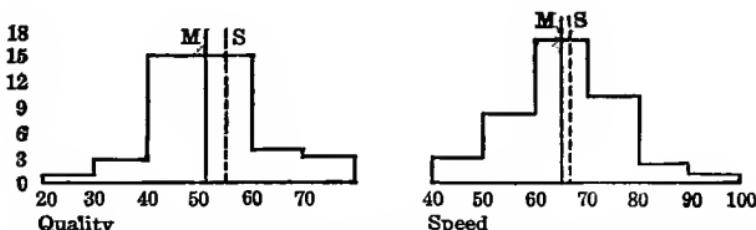


FIG. 13. SHOWING THE DISTRIBUTION OF SCORES IN HANDWRITING OF A FIFTH-GRADE CLASS

The line M indicates the median score for the class, the line S the standard for the class.

Of the conditions revealed in these figures, four types are significant for the teacher. These are low class-medians, a wide range of scores, low scores for individual pupils, and high scores for individual pupils. Low class-medians and low individual scores refer to median and individual scores which are lower than the proposed standards for the grades under consideration. High individual scores are probably not significant until they are higher than the standard for the eighth grade. A wide range of scores means that the pupils have not all profited equally by the instruction received. When the range is very wide it means that some pupils are either defective or suffering from neglect.

The scores for the third and fourth grades which are given in Figs. 11 and 12 show that the class-medians are low with the exception of the median for speed in Fig. 12. Assuming that the pupils should make a fairly steady development in their skill in handwriting from the second through the eighth grade, these low medians mean that the pupils in these classes are a year or more behind in their development. Either they must make very rapid progress in some one or more years, or they will leave the elementary school inadequately equipped in handwriting. When the class-medians are as near to the standards as those shown in the graph for the fifth-grade class, it is probable that the class can, by a little extra effort, reach the standard in a short time.

The graph for the scores in speed for the third-grade class, and both graphs for the fourth-grade class, show wide ranges of scores. The range of the speed scores for the fourth grade is obviously most unsatisfactory. In this case more

than a third of the class have scores as high as the standard for the seventh grade, and nearly as many have scores as low as the second-grade standard. This evidence suggests that the instruction in handwriting has had practically no effect on the speed of handwriting of some pupils, while other pupils in the class have learned to write at more than standard speed. If the instruction is given with a definite aim in view and is effective, the class will not show such distribution as that of the fourth-grade class. In such cases as those revealed by wide distributions of scores the examination of individual scores will suggest the remedy.

When individual scores are examined several classes of individuals will be discovered. Some will have scores up to the standards in both speed and quality. These individuals need give the teacher no concern, unless there may be some who have very high scores in both speed and quality. These may be excused from further drill in handwriting if this time can be used for other activities.¹ The remaining pupils are those having low scores in speed and high scores in quality, those having low scores in quality and high scores in speed, and those having low scores in both speed and quality. All of the pupils included under these three heads need special attention in the form of diagnostic measurement, as described above, and such instruction as is required to meet the needs revealed by this diagnosis.

Successive measurements to reveal progress. Successive measurements will reveal the progress which has been made by a class during a given period. Table XXIV shows the prog-

¹ See Freeman, F. N. "Handwriting"; in *Sixteenth Yearbook of the National Society for the Study of Education*, part 1.

ress made by each of five classes in one city system. Speed is given in letters per minute, and quality in scores on the Ayres Scale. These classes were measured in September, in November, and in January. The gain is shown at the bottom of each column. The fifth grade shows a type of progress in which there is a considerable gain in both speed and quality during each interval, the last scores being approximately up to the standards. The sixth grade shows another type of gain in which the first interval gave a considerable gain in quality with but little gain in speed. The second interval showed a loss in quality but a marked progress in speed.

TABLE XXIV. PROGRESS IN HANDWRITING IN ONE CITY

	Grades									
	II		III		IV		V		VI	
	Speed	Quality	Speed	Quality	Speed	Quality	Speed	Quality	Speed	Quality
September.....	8	38	10	37	38	35	40	37	56	36
November.....	38	39	34	40	46	42	56	44	60	48
January.....	46	38	34	43	64	44	66	54	84	47
Amount of gain —										
September to November.....	30	1	24	3	8	7	16	7	4	12
November to January.....	8	-1	0	3	18	2	10	10	24	-1
Total gain ...	38	0	24	6	26	9	26	17	28	11

Meeting the situation revealed. The measurement of the ability of pupils to write reveals situations which demand that the teacher be resourceful in finding methods and devices which will remedy the shortcomings which have been shown

to exist. These methods and devices should be selected in the light of facts which have been established by investigations of the learning process,¹ as it occurs in learning to write.

Systems of penmanship. There are not sufficient data from comparative studies of different penmanship systems to establish any single system as superior to others in its effectiveness to secure results in terms of speed and quality of handwriting.

Movement in handwriting. Graves² describes three kinds of movement: first, finger movement; second, "arm movement" in which there is some movement of the fingers and considerable movement of the arms; and, third, free-arm movement, in which "the respective movements of the fingers and of the arm are proportionally equal in amount." Of these three types of movement Graves concludes that arm movement seems to give greater speed. Nutt³ does not find this to be so. Since Nutt secured a positive result of no correlation between speed and movement, and since his measuring devices and methods were more objective than those used by Graves, it seems safe to conclude that movement does not influence speed in writing for a short time. The apparent greater ease of production of arm or muscular

¹ No attempt is made to review or to criticize the material which appears in numerous manuals of handwriting. Much excellent material which appears in *The Teaching of Handwriting*, by Freeman, is not even mentioned, because of lack of space. The difficulty of confining this discussion to the actual facts discovered through measurement of handwriting will be apparent.

² Graves, S. Monroe. "A Study of Handwriting"; in *Journal of Educational Psychology*, vol. 'vii, p. 486. (October, 1916.)

³ Nutt, H. W. "Rhythm in Handwriting"; in *Elementary School Journal*, vol. 17, pp. 432-45.

movement may result in greater speed if speed is measured during a long period of writing.

Nutt also found that arm movement comes with age and motor development. None of the systems of penmanship were found to develop any appreciable amount of arm movement in younger children. Copy book methods and the teacher's emphasis on arm movement develop about the same degree of arm movement in ages ten to fourteen. Special supervisors secure more arm movement in children of these ages, and also in nine-year-olds. Well developed arm movement did not produce better quality than movements in which the arm was moved but little. Neither did well developed arm movement show greater speed. Neither does better arm movement result in an increase in rhythm. The child's natural rhythm of motion is an important factor in his learning to write.

Rhythm. The rhythmic quality of the movement increases with age, but has no relation to amount of arm movement or to the quality of the writing. Nutt found that speed of writing and rhythm increase together. That is, children who score high in rhythm also score high in speed, are older than the other children, but may not use arm movement or produce a better quality of handwriting than other children.

Speed. Both Nutt and Graves have shown that speed increases with age. Nutt shows that speed increases with an increase in the rhythmic character of the movement. An important factor in the production of speed of handwriting is that of hand position. Graves says that a free and easy or loose-handed position is most conducive to speed. There is some evidence that girls write more rapidly than boys.

Quality and speed. Several studies have sought for a relation between speed and quality of handwriting. In the Cleveland Survey¹ it was found that "in general speed and quality vary inversely. But there is a middle series of speeds and qualities where improvement in one does not seem to interfere with the other." That is, outside of the limits which are approximately those of the proposed standards, efforts to secure an unusual degree of quality will reduce the speed, and *vice versa*. Several investigations of adults' handwriting show that they tend to increase the speed and reduce the quality. A general view of the results bearing on this point shows that the children who write a good quality on the average write as rapidly as those who write a poorer quality. This seems to be due to the natural rhythm of the children. If this rhythm is forced or disturbed unduly the quality suffers. Thorndike's results indicate that causing a pupil to write more slowly than his normal rate did not improve the quality of the handwriting. Nutt showed that increase of rhythm tends to slur over the abrupt strokes, while an increase in speed tends to slur the difficult junctions of strokes. Freeman says the so-called muscular movement produces a firmness and evenness of line and a regularity of slant.

General laws of learning applied. The ability to write well is a habit, hence the laws of habit formation apply to the acquisition of this ability.

The *first* essential factor is a right start. The pupil must have a clear view of the habit to be acquired. This may mean a definite idea of the movement to be executed, or a

¹ Judd, C. H. *Measuring the Work of the Public Schools*, pp. 80-81.

picture of the letters or series of letters which are to be made. The start must be made with a strong initiative. Sometimes the pupil must be shocked into a desire to correct a fault of his handwriting.

The *second* essential is that of attentive repetitions. The repetitions or drills should be strongly motivated. All investigations of habit formation agree upon this point. The periods of practice are most efficient if not carried to the point of fatigue, hence, for the lower grades Freeman suggests frequent ten-minute periods of practice. In no grades should the periods be longer than twenty minutes.

The *third* step, as often stated, is, "Allow no exceptions to occur." If a pupil practices correct form in the penmanship class for ten minutes, and then uses poor form in a spelling class for the same length of time, the latter exercise will tend to cancel the effects of his practice in the penmanship class. In view of the frequent occurrence of such conditions as these, special periods, during a part or all of a term, may be set aside for intensive penmanship study and practice.

A *fourth* step is the repetition of the habit until it is well fixed. This means that the repetitions must extend beyond the point of apparent completion to permanent automatism. After this stage is reached incentives should be found which will raise the habit from the level of mere automatism to higher levels of skill.

Devices of remedial instruction. A few devices are given which are not usually found in the manuals, and which meet the specific needs revealed by measurements. Other devices may be found in the manuals issued by the representatives of the several systems of penmanship.

Increasing speed. When a pupil habitually writes slower than the standard it will be well to force this pupil to write at standard speed. The influence of the increased speed can then be observed. If the teacher uses music as an aid to rhythm, the faster time of the music may increase the rate of a pupil's handwriting, but to insure this, the speed of handwriting should be carefully measured by accurate timing and actual count of the letters. A dictated exercise will accomplish the result more surely, and with economy of time and effort for the teacher. For example, the sentence, "The quick brown fox jumps over the lazy dog" contains thirty-five letters.

8th-grade pupils should write this 11 times in 4 min.

7th	"	"	"	"	"	8	"	"	3	"	30 sec.
6th	"	"	"	"	"	6	"	"	3	"	
5th	"	"	"	"	"	5	"	"	2	"	45 "
4th	"	"	"	"	"	4	"	"	2	"	30 "
3d	"	"	"	"	"	3	"	"	2	"	10 "
2d	"	"	"	"	"	2	"	"	2	"	

The pupils should memorize the sentence and write it several times for practice and for spelling. The teacher should then time their writing, according to the table given above. Those who do not write the required number of letters in the allotted time should be told to write faster, until they have done the test successfully.

These time intervals are calculated to meet the requirements for speed as furnished by the Freeman standards. Another device which is even more convenient is represented by the following example. This is a dictation exercise for the sixth grade. The teacher should direct the class to be ready to write, then, watching the second hand of her watch,

until it is at 60, start to dictate. A little preliminary practice will make it easy to dictate the words so that they will be pronounced as indicated. For example, the teacher should be pronouncing the word "care" just before the second hand reaches the ten second mark, etc.

5 10 20 30
Do you take care to keep your teeth very clean, by wash-
ing them without failing every morning and after every
meal? This is very necessary both to preserve your teeth
a great while, and to save you a great deal of pain. (Stop.)

40 50 60
15 20 30 40 50
60 10 20

Developing rhythm. If this exercise reveals a serious sacrifice in the quality, or if the pupil's handwriting cannot be brought up to standard speed, we may consider that the pupil's rhythm has not developed to the place where it will sustain this speed. Since we do not know which is the primary factor, rhythm or speed, the best procedure would be to seek to develop both. Rhythm may be increased by the use of music. Careful attention to the securing of a free, well-relaxed hand position will aid in securing speed. Sometimes a careful analysis of letter forms will reveal that the student is forming some letters in a way that makes speed impossible. In such cases new forms of those letters should be taught. Sometimes the written work in classes other than the penmanship class places unusual demands on the pupil's powers. Throughout the lower grades written work, which places demands on the pupil's speed which are different from those of the writing lessons, should be avoided.

In seeking to improve the quality of a pupil's handwriting,

analysis should be made by means of the Freeman Scale or the Gray Score Card. After these analyses sometimes it is well to call the pupil's attention to a single letter or combination, and make this vivid by means of the warnings or suggestions which are most effective. When a single fault is corrected another may be attacked, until the pupil acquires this power to correct his faults.

Motivating practice. A number of devices and plans have been proposed for the motivation of practice in correcting faults in quality of handwriting. Wilson¹ gives the result of an interesting experiment in which the Thorndike Scale was used in such a way that the students could follow their own progress in handwriting. In this case each student was competing with his own record. Several have constructed scales from the specimens collected in a school or class. These scales may be constructed by rating the specimens with any one or more of the scales described. Superintendent Bliss, of the Montclair, New Jersey, schools is quoted by Wilson as follows: "A scale made from the writing of pupils makes a stronger appeal than either the Ayres or Thorndike Scales."

Charters² recommends a "writing hospital" to which the poor writers are sent until they are properly convalescent. This hospital is a special penmanship class. Stone³ has a plan which puts all the pupils of a school in four groups for their writing lessons. These are groups 1, 2, 3, and the ex-

¹ Wilson and Wilson, *The Motivation of School Work*, p. 187. (Houghton Mifflin Company, 1916.)

² Charters, W. W. *Teaching the Common Branches*. (Houghton Mifflin Company, 1913.)

³ Stone, C. R. "Motivation of the Formal Writing Lesson Through a Special Classification of Pupils for Writing"; in *School and Home Education*, June, 1915.

cused groups. The special feature of this plan is that at stated intervals members of a lower group are allowed to challenge members of a higher group, and a contest for the coveted place ensues.

Many special devices for motivation are in use. Pupils write letters ordering supplies for the school, or they write invitations to school parties, pageants, etc. Some pupils write letters for the teacher or principal.

Reasons for using handwriting scales. Even when measures of handwriting are not accurate they force the teacher to give attention to the specific faults and needs of the pupils. This measurement creates a critical and scientific attitude in the teacher toward the outcomes of instruction. This attitude tends to remove the attention from personal bias and feeling to an objective consideration of the results secured. Measurement of handwriting also banishes the old false standards represented by the perfect specimens which were produced from an engraved plate. In their stead are proposed some standards which are within the reach of a majority of the pupils. Thus many children can know the joy which comes from achieving something recognized to be of value. In addition to these values measurement is destined to become scientifically accurate and thus furnish a valid basis for instruction.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. A teacher may judge the handwriting of her class by watching the pupils while they write or by examining the specimens which they have written. Which is the better method if the purpose is to make comparisons of classes? Which is better for discovering the handwriting defects of individual pupils? What factors would you keep in

mind in watching children while they write? What factors in the other method?

2. Ask a class to write the three sentences from Lincoln's Gettysburg address. Direct them to start together and write as *rapidly* as they can for one minute. At the end of one minute stop them and direct them to record their speed of handwriting. Then ask them to begin again and write for one minute writing as *well* as they can. If you wish to eliminate practice effects, repeat the experiment again reversing the order of the directions. Note the different effects due to the nature of the directions.
3. Use the dictation exercises given on page 185. First dictate at the standard rate for the class. Next dictate at the standard rate for a class two grades lower. Then dictate at the standard rate for a class two grades higher. Examine the specimens using a standardized scale and note the effects of the different rates of writing on the quality.
4. Try several different selections for copy when collecting specimens and determine the selection which your pupils will write most rapidly.
5. In what situations would you use the Ayres Scale? Thorndike Scale? Johnson-Stone Scale? Freeman Scale? Gray score card?
6. What factor in handwriting is of most importance according to the Gray score card? What factor is of least importance? Why are these factors so rated?
7. Select ten or preferably one hundred specimens of handwriting and rate them every day for several days by means of the scale you have. Keep the record of your day's rating but do not use them to help you in making future ratings. After several ratings note the consistency of your ratings.
8. Use the Gray Score Card (or Freeman Scale) in scoring the poorer specimens of handwriting. Prescribe the drills you would use in correcting these defects. Compare this with the recommendations of other teachers or students. Try your prescription on the pupils concerned if possible.
9. For what purpose would you use the dictation exercises?
10. Use the Gray Score Card by filling it with scores secured from use of the Freeman Scale, whenever they apply.
11. Select a defect of letter formation frequently found in a pupil's handwriting. Direct the pupil's attention to this defect and challenge him to correct it. Direct that a record be taken as follows: If the defect were found in letter "a" instruct the pupil to count the number of such errors to be found in fifty consecutive "a"s as they occur in his handwriting written prior to the time you pointed out the defect. After a period of practice, direct the pupil to make another counting from his handwriting written at some period other than the writing period.

BIBLIOGRAPHY

Only the most important references are given here. Additional references will be found in the footnotes of the chapter.

I. SCALES AND REFERENCES

1. *Ayres's Measuring Scale for Handwriting*. Copies may be obtained from Division of Education, Russell Sage Foundation, New York City.

REFERENCES: Ayres, L. P. *A Scale for Measuring the Quality of Handwriting of Children*.

Second Annual Report of Bureau of Educational Measurements and Standards, 1915-16. (Kansas State Normal School.)

Ashbaugh, E. J. *Handwriting of Iowa School Children*. (University of Iowa, Extension Division, Bulletin no. 15.)

King, Irving, and Johnson, Harry. "The Writing Abilities of the Elementary and Grammar School Pupils of a City School System Measured by the Ayres Scale"; in *Journal of Educational Psychology*, vol. 3, pp. 514-20. (November, 1912.)

2. *Ayres's Scale for Measuring the Quality of Handwriting of Adults*. Copies may be obtained from the Division of Education, Russell Sage Foundation, New York City.

REFERENCE: Ayres, L. P. *A Scale for Measuring the Quality of Handwriting of Adults*. (Bulletin E 138, Division of Education, Russell Sage Foundation, New York City.)

3. *Freeman's Handwriting Scale*. Copies may be obtained from Houghton Mifflin Company.

REFERENCES: Freeman, F. N. *The Teaching of Handwriting*.

Freeman, F. N. "An Analytical Scale for Handwriting"; in *Elementary School Journal*, January, 1915.

4. *Gray's Score Card for the Measurement of Handwriting*. Copies may be obtained from C. T. Gray, University of Texas, Austin, Texas.

REFERENCE: *A Score Card for the Measurement of Handwriting*. (Bulletin no. 37, University of Texas.)

5. *Johnson-Stone Scale*. Copies not obtainable.

REFERENCE: Johnson, George L., and Stone, C. R. "Measuring the Quality of Handwriting"; in *Elementary School Journal*, February, 1916.

6. *Thorndike's Scale for Measuring the Handwriting of Children in Grades 5 to 8*. Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCES: Thorndike, E. L. "Handwriting"; in *Teachers College Record*, March, 1910.

Thorndike, E. L. "Teachers' Estimates of the Quality of Specimens of Handwriting"; in *Teachers College Record*, vol. 15, no. 5. (November, 1914.)

7. Ayres's "Gettysburg Edition" copies may be obtained from the Division of Education, Russell Sage Foundation, New York City.

II. GENERAL REFERENCES

Breed, F. S., and Down, E. F. "Measuring and Standardizing Handwriting of a School System"; in *Elementary School Journal*, vol. 17, no. 7, p. 470. (March, 1917.)

Freeman, Frank N. "Handwriting Tests for Use in School Surveys"; in *Elementary School Journal*, vol. 16, pp. 299-301. (February, 1916.)

Freeman, Frank N. "Handwriting"; in *The Fourteenth Yearbook of the National Society for the Study of Education*, part I. University of Chicago Press, 1915, chap. v. Also in the *Sixteenth Yearbook of the National Society for the Study of Education*, part I. (1917.)

Graves, S. Monroe. "A Study in Handwriting"; in *Journal of Educational Psychology*, vol. 7, pp. 483-94. (October, 1916.)

Johnson, Joseph Henry. *A Comparison of the Ayres and Thorndike Handwriting Scales*. Containing a table of equivalent values in the two scales. (North Carolina High School bulletin, vol. 7, pp. 170-73. October, 1916.)

Manuel, Herschel T. "Studies in Handwriting"; in *School and Society*, vol. 10, no. 116. (March 17, 1917.) Gives the frequency of letters and elements of letters in writing as shown by the Ayres Spelling List. Suggests certain advantageous changes in forms of letters.

Nutt, H. W. "Rhythm in Handwriting"; in *Elementary School Journal*, vol. 17, pp. 432-45.

CHAPTER VI

LANGUAGE

I. THE PROBLEM OF MEASUREMENT IN LANGUAGE

One of measuring specific habits. Language functions in the communication of those things which are commonly called ideas and feelings by means of words. The choice and arrangement of the words give language its form. In written language spelling and handwriting contribute additional elements of form. The ideas and feelings which language communicates may be described as its content.

The rules of grammar definitely prescribe many items of form. For example, a verb must agree with its subject in number and person; pronouns are inflected for person, case, gender, and number; verbs are inflected for mode, tense, person, and number; certain words must be capitalized. A pupil's control of those items of form which are definite and which occur frequently must be reduced to the plane of habit or automatic functioning, so that his attention may be focused upon the content which he is attempting to express. For these abilities the problem of measurement is the problem of measuring specific habits, and in this respect it is similar to the problem of measurement in the subjects treated in the preceding chapters.

Rhetoric treats of the choice of words and the structure of sentences and paragraphs, but it does not prescribe definite objective standards for them. The quality of these features of form is determined by the effect of the lan-

guage upon the reader, and this effect is not the same for all readers. However, rhetoric does furnish certain general principles which are useful to the pupil in guiding his construction of a form which will attain his purpose. Here the problem of measurement is different and more difficult. The functioning of the principles cannot be reduced to the plane of habit, because it is necessary that they function in a variety of new situations.

The content of language is subtle and is not objective, except as it is given a form. It depends upon the vividness and the organization of ideas, and upon the wealth of associations which give the central ideas their setting. These features of content are expressed through the choice of words and the structure of sentences and paragraphs. In this way content and form are so intimately connected that aside from the features of form which are specified by the rules of grammar, any attempt to measure one is made difficult by the presence of the other.

The instruments for measuring ability in language may be divided into two classes. The composition scales and Trabue's completion-test language scales are instruments for general measurement of language ability. The grammar scales and the copying test measure specific features of form.

II. THE MEASUREMENT OF ABILITY IN ENGLISH COMPOSITION

For measuring compositions written by school-children scales similar in plan to the handwriting scales have been devised. The first of these scales is the Hillegas Composition Scale. A revision of this scale has been published with

the title, *Preliminary Extension of the Hillegas Scale*, by E. L. Thorndike. Other scales are:¹ The Harvard-Newton Composition Scale, devised by F. W. Ballou; A Scale for Measuring the General Merit of English Composition in the Sixth Grade, devised by F. S. Breed and F. W. Frostic; A Scale for Measuring Written English Composition, devised by M. H. Willing; and the Nassau County Supplement to the Hillegas Scale, devised by M. R. Trabue. Each of these scales consists of compositions arranged in the order of their merit. The relative merit of the compositions was determined by means of careful statistical studies.²

1. The Hillegas Scale.

This consists of ten compositions

¹ Standard Tests in English, by S. A. Courtis, are not included in this list. These tests were a group of tests in reading and language and proved to be so cumbersome to use that their publication has been discontinued.

² No attempt is made in this chapter to give the methods employed in deriving these scales nor to summarize the criticisms which have been made upon the methods. For these the reader is referred to the following: —

Hillegas, M. B. "A Scale for the Measurement of Quality in English Composition for Young People"; in *Teachers College Record*, vol. 13, no. 4. (September, 1912.)

Ballou, F. W. "Scales for the Measurement of English Compositions"; in *Harvard-Newton Bulletin*, no. 2. (Harvard University.)

Kayfetz, Isidore. "A Critical Study of the Hillegas Composition Scale"; in *Pedagogical Seminary*, vol. 21, pp. 559-77. (December, 1915.)

Kayfetz, Isidore. "A Critical Study of the Harvard-Newton Composition Scales"; in *Pedagogical Seminary*, vol. 23, pp. 325-47. (September, 1916.)

Brownell, Baker. "A Test of the Ballou Scale of English Composition"; in *School and Society*, vol. 4, pp. 938-42.

Breed, F. S., and Frostic, F. W. "A Scale for Measuring the General Merit of English Composition"; in *Elementary School Journal*, vol. 17, pp. 307-25.

Willing, M. H. *Measurement of Written English Composition in the Public Elementary Schools of Denver, Colorado*. (Master's Thesis, Chicago.) See also *Report of the School Survey of School District Number One in the City and County of Denver*, part II, p. 59.

Trabue, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. 18, p. 51. (January, 1917.)

ranging from an artificial production whose scale value is zero to the tenth composition whose scale value is 9.3. Three of the ten compositions are artificial productions, five were written by high-school pupils, and the remaining two by college freshmen. No two were written on the same topic and they vary greatly in length and type. Each degree of merit is represented by only one composition. In the Thorndike Extension of the Hillegas Scale only a few of the compositions of the original scale have been used and several compositions are given for each degree of merit in the middle of the scale. Twenty-nine compositions represent fifteen degrees of merit within approximately the same range as the original scale. This makes a more finely divided scale than the original one.

2. The Harvard-Newton Scale. The Harvard-Newton Composition Scale consists of four separate scales, one for each form of discourse; argumentation, description, exposition, and narration. Each of the scales consists of six compositions written by eighth-grade pupils and arranged in order of merit as determined by the marks assigned by teachers rating them as eighth-grade compositions. For each composition there is given a statement of the most significant merits and defects.

3. The Breed and Frostic Scale. The compositions used by Breed and Frostic in deriving their scale were written by sixth-grade pupils under uniform conditions. A part of a story called "The Picnic" was read to the class and they were given twenty minutes to complete it. The method of selecting compositions for the scale and determining scale values was similar to that employed by Hillegas.

4. Willing's Scale. Willing used compositions written by pupils in grades four to eight on the topic, "An Exciting Experience." Several particular exciting experiences were suggested, and twenty minutes were allowed for writing. In determining the compositions to be used for the scale, "all errors in spelling, punctuation, capitalization and grammar were counted and corrected." The relative merit of the corrected compositions was determined and those compositions were selected for the scale which had the same rank in "story value" and frequency of errors. The scale is reproduced on pages 206-10.

5. The Nassau County Supplement. The Nassau County Supplement to the Hillegas Scale consists of nine compositions, seven of which were written by elementary school pupils on the topic, "What I should like to do next Saturday." The compositions of the scale were carefully selected and evaluated by an elaborate method which cannot be even sketched here.

Reliability of measurements. These scales are to be used in the same way as the handwriting scales. In measuring the ability of pupils in the field of English composition, the first step is to secure compositions written under defined conditions. After the compositions are obtained the merit of each is measured by comparing it with the compositions which make up the scale used and its degree of merit is that of the scale composition which it most closely resembles. Assuming that the degree of merit of the scale compositions has been accurately determined, the accuracy of the measures obtained depends upon the reliability of the comparison.

The accuracy of measurements made by using the Hillegas Scale has been investigated by having the same compositions rated by a group of teachers, first, by the usual method and second, by using the scale. By means of a number of such investigations the conclusion has been reached that "the variability is somewhat greater with the scale than without it."¹ However, in the investigations reported the teachers using the scale were untrained in its use. Furthermore, as in the case of handwriting practically no attention has been given to determining the best methods of using a composition scale. Because of these two facts the conclusions reached in the studies just referred to must be qualified. Thorndike has asserted that errors in using the scale will diminish with practice² and with sufficient practice they will be smaller than the errors now made by teachers in grading paragraph writing for general merit. Trabue states that "In spite of all criticisms of and objections to the Hillegas Scale, the fact remains that it is one of the most useful measuring instruments in the whole field of education."³

An investigation of the reliability of the measurements made with the Harvard-Newton Scale has been reported.⁴

¹ Kelly, F. J. *Teachers' Marks*, p. 134.

² Thorndike, E. L. "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition"; in *English Journal*, vol. 2, p. 551.

³ Trabue, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. 18, p. 51.

⁴ "Second Annual Conference on Educational Measurements"; Indiana University Bulletin, vol. 13, no. 11, pp. 115-22. Also Hudelson, Earl. "Some Achievements in the Establishment of a Standard for the Measurement of English Composition in the Bloomington, Indiana, Schools"; in *English Journal*, November, 1916.

Compositions written by 386 pupils in grades 7, 8, and 9 were used. The variability of the marks assigned by means of the scale was slightly less than when the scale was not used.

The other scales have appeared only recently and no study of their reliability has been reported. Since the Thorndike Extension of the Hillegas Scale is more finely divided and the degrees of merit are represented by more than one composition it is possible that it will yield more reliable measures than the original scale. The scales devised by Trabue, Willing, and Breed and Frostic consist of compositions written on the same topic and under known conditions. When used to measure the merit of compositions written on the same topic and under the same conditions it seems probable that any one of these scales would yield a more reliable measure of a pupil's ability in written expression. At least the procedure is more scientific.

Use of the scales. The final test of the reliability of the measures obtained by using these scales must be based upon their use by teachers who have been trained in using them. The results of studies based on handwriting scales suggest that practice in the use of a composition scale will materially increase the reliability of measures. In the absence of a scientifically determined plan of training, the plan given for handwriting (page 165) may be used. Even if it is shown that composition scales have no value as instruments for measuring the merit of compositions, it does not follow that the scales are without value to the teacher. The scales represent degrees of merit in English compositions, and thus assist the teacher in setting before her pupils the standards toward which they should strive.

Particularly is this true of the Harvard-Newton Scale. By pointing out the merits and defects of each composition the pupil is given an objective statement of what the teacher expects of him.

Directions for using the Hillegas Scale. The Hillegas Scale has been used in the surveys of Butte, Montana, and Salt Lake City, Utah. At Salt Lake City the following directions were followed for securing compositions written by pupils:—

1. Each teacher is requested to ask her children to write a composition for her on the following theme:—

“Suppose that you have twenty dollars, which you have been given to spend. You have five friends, and you decide to spend it in such a manner as will give the most pleasure to each. Tell what you would do or buy for each friend. The amount spent for each friend need not be the same, but the total for the five must be twenty dollars.”

2. The composition should be written with pen and ink on the regular writing paper.

3. After the children are ready for writing, read the subject to them, give them a minute or two to ask any questions, and as soon as you are sure that the children understand what they are to do, start them at writing.

4. When the children have finished collect the papers, fasten those for each class together with a clip, and send to the office of the school principal.

Hillegas Scale Scores. Similar directions were used at Butte. In both instances the compositions were rated by the teachers, using the Hillegas Scale, but no teacher rated the compositions written by her own pupils. The median scores for the two cities and the standards proposed by Trabue¹ are given in the following table:—

¹ Trabue, M. R. “Supplementing the Hillegas Scale”; in *Teachers College Record*, vol. 18, p. 51. (January, 1917.)

<i>Grade</i>	<i>Salt Lake City*</i>	<i>Butte †</i>	<i>Trabue : Median score</i>	<i>Trabue : Score above which three fourths of the pupils should rank</i>
XII.....	7.2	6.7
XI.....	6.9	6.4
X.....	6.5	6.0
IX.....	6.01	5.5
VIII.....	5.4	4.11	5.5	5.0
VII.....	4.4	3.75	5.0	4.5
VI.....	3.8	3.40	4.5	4.0
V.....	3.1	2.87	4.0	3.5
IV.....	2.9	2.34	3.5	3.0

* *Report of a Survey of the School System of Salt Lake City, Utah*, p. 145. (1915.) A revised edition has been published (1916) by The World Book Company, under the title, *School Organization and Administration*, by E. P. Cubberley.

† *Report of a Survey of the School System of Butte, Montana*, p. 76. (1914.) A revised edition has been published (1916) by the same firm under the title, *Some Problems in City School Administration*, by G. D. Strayer. See p. 167.

While the median scores for Salt Lake City show a rather even rate of progress from grade to grade, and are distinctly higher than the Butte median scores, a charting of the distribution of the scores attained in each of the grades (see Fig. 14) reveals a distribution all along the scale. Many fourth-grade children wrote better than many eighth-grade children, and so for all of the grades. Such charting clearly reveals the individual and class problems with which teachers and supervisors have to deal. Fig. 14 shows not only the range of distribution of the scores obtained by the pupils of each grade, but also the per cent of pupils in each who attained each of the possible scores and the median score for each grade.

Directions for using the Harvard-Newton Scale. When

Hillegas
 Scale Scores .0 1.83 2.60 3.39 4.14 5.85 6.75 7.72 8.38 9.37
 Numbered Scores 0 1 2 3 4 5 6 7 8 9

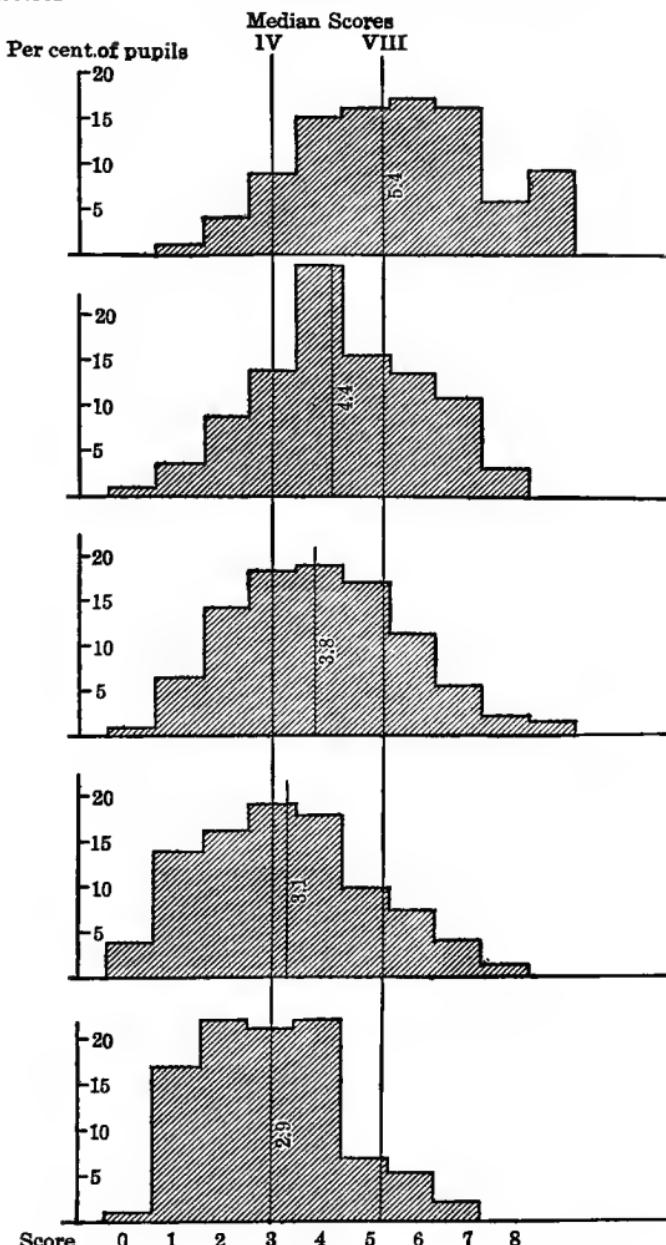


FIG. 14. RESULTS OF THE COMPOSITION TEST IN SALT LAKE CITY

using the Harvard-Newton Scale the compositions should be written under approximately the same conditions as were followed in obtaining the compositions for the scale. The following directions prepared on this basis have been widely used.¹

HOW TO OBTAIN COMPOSITIONS

The compositions should be written as a part of the regular school work. Give two periods of forty minutes each for this work — one for preparation and writing and one for correcting and copying. Use one or two days, whichever is more convenient, but if two days are taken, collect the papers at the end of the first period and distribute them the next day. The pupils may use the dictionary and other reference books for preparation, but not while writing, correcting or copying. The compositions should be written in ink. The teacher should answer no questions after the writing begins.

SELECTING A SUBJECT

The teacher should suggest a subject from the following list. Then permit each pupil to choose his subject, either one of those suggested to him or one proposed by him, except that it must not be a subject upon which he has written recently.

BOYS

Outdoors :

- The best way to catch rats.
- How to build a shelter for the night.
- The hired man.
- When I helped father harvest.
- Should a town boy own a dog?
- When the horse ran away.

Mechanics :

- How to run a Ford.
- Is a Ford better than a Saxon?
- Our mowing machine.
- How to dam the creek.
- How to make a rabbit trap.

Games :

- My first baseball game.
- Should football be abolished?
- Is _____ better than _____? (Fill in the names of two authors of juvenile stories.)
- Should a boy be made to go to school?
- The use of keeping pet rabbits.
- Is it right to catch fish with a hook?
- Should a boy play marbles for keeps?

¹ Prepared by the Bureau of Educational Measurements and Standards, Kansas State Normal School.

GIRLS

Home:

My room.	<i>Social:</i>
How to get the best breakfast I know.	My new dress.
When the lamp tipped over.	How to eat soup.
Which is better in the home, music or art?	My first party.
Ought boys and girls take piano lessons?	When Mr. ——'s house burned.
	The worst boy I ever knew.

School:

When a bird came into the schoolroom.
 How I killed giant Laziness.
 Is Latin more useful than science?
 The inside of my desk.
 The boy across from me.

BOYS AND GIRLS

When I become a voter.	<i>Is a college education worth while?</i>
The work I like best.	Teacher.

After a subject has been selected by each pupil, direct him to write a composition "about one hundred fifty words in length, not over two hundred words." When the composition is finished ask the pupil to write his name, grade, and the date upon the back.

The Harvard-Newton Scale Scores. A number of copies of the Harvard-Newton Scale have been placed in the hands of teachers, but no report of its use with a large number of compositions is available.

In the investigation of the composition work in the schools of Bloomington, Indiana (see footnote 4, on page 197), using this scale, the compositions were limited to description. Seven topics were given: (1) Some Person in Bloomington; (2) Grandfather; (3) An Old-Fashioned House; (4) A Picture; (5) A Public Building in Bloomington; (6) A Body of Water; (7) A Wreck. Thirty minutes were given for actual composition of the first draft, ten minutes for preliminary explanation, and thirty minutes for writing. Later the pupils were allowed to correct and

rewrite for thirty minutes more. Each composition was rated by three teachers, and, based on the average of these three ratings. Huddleston obtained the following median scores:

<i>Grade</i>	<i>Number of compositions</i>	<i>Median score</i>
IXa.....	61	68
IXb.....	57	68
VIIIa.....	61	69
VIIIb.....	68	66
VIIa.....	72	64
VIIb.....	67	60

Directions for using the Willing Scale. In using the Willing Scale these directions should be followed.¹

The teacher should write on the blackboard these topics:²

An Exciting Experience

- A Storm
- An Accident
- An Errand at Night
- A Wonderful Story
- An Unexpected Meeting
- In the Woods
- In the Mountains
- On the Ice
- On the Water
- A Runaway

The teacher should then say to the pupils: "I want you to write me a story. It is to be a story about some exciting experience that you have had, about something very in-

¹ These directions are based upon those followed by Willing in securing compositions for his scale. The author is indebted to Mr. Willing for a copy of his thesis and for permission to print his scale.

² It is probably better to furnish each pupil with a printed list of the topics.

teresting that has happened to you. If nothing of the sort has ever happened to you, then tell me of an exciting experience some one you know has had. You may even make up a story of this kind, if you have to, though I believe you will do better, on the whole, with a real one. I am going to give you about twenty minutes in which to write. You are to write on both sides of the paper, to do all the work yourselves, and to ask no questions at all after you begin. You may make whatever corrections you wish between the lines. There will be no time to rewrite your story.

"I have written the general subject on the board together with some suggestions. You do not have to write on any of these topics unless you want to; they are merely to help out in case you cannot think of an exciting experience yourself. You may begin now as soon as you wish."

Allow opportunity for asking questions and make an effort to put the children at ease. Allow full twenty minutes for the actual writing. At the end of this time say to the pupils:—

"You are to have four or five minutes in which to finish your stories, make corrections and count the number of words written. Write this number at the end of your story. Also write your name and school grade." At the end of five minutes collect the papers.

The method of using the scale in the Denver Survey is given by Willing as follows:—

The method of using the scale which was developed from almost the start was as follows: A composition was read carefully, with attention to both rhetorical and, what we have been calling, formal elements. As the reading progressed, there was a conscious effort to place it on the scale, so that by the end of the reading its

fate had frequently been decided. But in many, many cases the discrepancy between the story value and the form value was such as to dictate an adjustment or compromise of some sort. In these cases it was the writer's habit to assign the story value first and then try to locate the formal worth. A compromise between the two had then to be assigned as the mark of the paper. Thus a composition grading very low as a story might by excellence in formal matters achieve as high a mark as 60; and, vice versa, a composition of high story value but low formal quality might be marked as far down as 40. But no paper was marked above 70 which did not have both good story value and technical excellence to commend it; nor was a paper marked below 40 which did not lack both of these qualities. It is not possible to state exactly the relative emphasis that was placed on story value and form value, but the effort was made, within the limits just mentioned, to keep the two approximately equal. The interpolated marks, 25, 35, etc., were used in grading.

The use of this scale has been shown to be quite consistent when one judge rates all of the papers. It still remains to show how consistent the ratings will be when several judges are employed.

The Willing Scale Scores. For the Denver Survey the following median scores were obtained:—

4th A	31.5
5th A	43.4
6th A	50.9
7th A	60.2
8th A	63.4

WILLING SCALE FOR MEASURING WRITTEN COMPOSITION

(The values: 90, 80, 70, 60, 50, 40, 30, and 20 given the respective samples are arbitrary and merely for practical convenience. 20 means 15 to 24.9, 30 means 25 to 34.9, etc.)

A—90

The most exciting experience of my life happened when I was but five years of age. I was riding my tricycle on the top of our

high terrace. Beside the curbing below, stood a vegetable wagon and a horse. Suddenly I got too near the top of the terrace. The front wheel of my tricycle slipped over and down I went, lickety-split, under the horse standing by the curbing. I had quite a high tricycle and the handle-bars scraped the horse's stomach, making him kick and plunge in a very alarming manner. I was directly under him during this, but finally rolled over out of his way and scrambled up. I looked at my hands! Most of the first finger and part of the thumb of my left hand were missing. The horse had stepped on them. I had endured no sensation of pain before this, but now my mangled hand began to hurt terribly. I was hurried to the hospital and operated on, and now you would hardly notice one of my fingers is missing. I certainly have good cause to congratulate myself on my good fortune in escaping with as little injury to myself as I did, for I might have been terribly mangled in my head or body.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 0.

B — 80

Near our ranch in Fort Logan there was a chicken ranch. One day my sister and I went up to the chicken ranch on our horses. Coming back there was a road leading from our house to the main road and along this road were half rotted stumps. On every one of these stumps what do you think we saw. We saw snakes! snakes! snakes! I suppose these snakes were shedding their skins, they were of every color, shape, and size. But when sister and I saw these snakes we whipped our horses into a gallop and away we went just as hard as we could go. When we got to the house we went in and mamma couldn't get us out of the house that day. I was so scared that I believe I dreamed about snakes for a month.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 5.

C — 70

When I was in Michegan I had an exciting thing happen or rather saw it, it was when the big steamship plying between Chicago and Muskegon was sunk about 7 o'clock in the evening.

It caught on fire with a load of cattle and products from the market on board, one of the lifeboats carrying some of the few people who were on board landed at our pier. The "Whaleback" steamer which goes between Chicago and Muskegon was two hours later in coming than the freighter and was stopped to clear up the wreckage. all of the cattle and products and an immense cargo of coal were lost, but there were only two people lost. the ship tried hard to get to port with her cargo but, could not reach it. The next morning we found planks, and parts of the wreck on the beach. Our cottage was at the top of a cliff and it was just one hundred feet to the lake from our cottage, we had a beautiful view, and the sight of the fire on the horizon was a beautiful sight (though it was pitiful).

Number of mistakes in spelling, punctuation, and syntax per hundred words — 8.

D — 60

One time when mother, some girl friends and myself were staying up in the mountains. An awful storm came up. At the we were way up the mountain. The lightning flashed and the thunder roared. We were very frightened for the cabin we were staying at was at the foot of the mountain. We did n't have our coats with us for it was very warm when we started. There were a few pine trees near us so we ran under them. They did n't do much good for the rain came down in torrents. The rain came down so hard that it uprooted one of the trees. Finely it began to slack a little, So we thought we would try and go back. About half way down the mountain was a little hut. We started and when got about half way down it began to rain all the harder. We did n't know what to do for this time there was n't any trees to get under. We decided to go on for the nearest shelter was the hut. Finely we got there cold and wet to the skin.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 11.

E — 50

One time mother and father were going to take sister and I for a long ride thanksgiving, We had to go 60 miles to get there, When

sister and I herd about it we were very glad. It was a very cold trip. We four all went in a one seated automobile. Dady drove and mother held me and sister sat on the top the top was down. Mother could not hold sister for she was two heavy. When we got there they had a hot fire ready for us and a goose dinner. We were there over night. In the morning it was hot out. This was on a farm. Sister and I got to go horse-back riding. It was lots of funs. They had children. The children were very nice. Our trip home was very cold. When we got home it had snod.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 14.

F — 40

My antie had her barn trown down last week and had all her chickens killed from the storm. Whitch happened at twelve o'clock at night. She had 30 chickens and one horse the horse was saved he ran over to our house and claped on the dor whit his feet. When he saw him my father took him in the barn where he slepped the night with our horse. When our antie told us about the accident we were very sorry the next night all my anties things were frozen. The storm blew terrible the next morning and I could not go to school so I had to stay home the whole week.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 17.

C — 30

The other day when I was rideing on our horse the engion was comeing and he got frightened so he through me down and I broke my hand.

And the next thing I done was I went to the doctor and he put some bandage on it and he told me to come the next day so I came the next day and he toke the bandage off and he look at it and then it was better.

Number of mistakes in spelling, punctuation, and syntax per hundred words — 23.

H — 20

Deron the summer I got kicked and sprain my arm. And I was in bed of wheeks And it happing up to Washtion Park I was going to catch some fish. And I was so happy when I got the banged of I will nevery try that stunt againg.

Number of mistakes in spelling, punctuation, and syntax per hundred words --- 30.

III. THE MEASUREMENT OF LANGUAGE ABILITY BY COMPLETION TESTS

The Trabue Completion-Test Language Scales. Trabue has devised a series of Completion-Test Language Scales for the general measurement of language ability. Of these scales he says: —

No attempt has been made to define "language" in any strict sense, and it is entirely possible that some persons may be able to speak the English language and perhaps to write it fairly well without being able to make a very high score on these scales. It may also happen that some individuals will be found who score well on these language scales and are yet unable to quote a single rule of English grammar. On the whole, however, it will be found that ability to complete these sentences successfully is very closely related to what is usually called "language ability."¹

Each scale consists of sentences from which one or more words have been omitted. The position of the omitted words is indicated by a blank. The pupil being tested is to write in the missing words. The relative difficulty of the sentences has been carefully determined and they have been arranged in order of difficulty. Directions for giving the tests, scoring the papers and tabulating the results are given

¹ Trabue, M. R. *Completion-Test Language Scales*, p. 1. Teachers College Contributions to Education, No. 77. 1916.

by Trabue on page 20 of his monograph. Scale B is reproduced here to illustrate this type of test.

TRABUE COMPLETION-TEST LANGUAGE SCALE B

Sentence number	Value	
1	.96	We like good boys girls.
6	1.98	The is barking at the cat.
8	2.94	The stars and the will shine to-night.
22	4.26	Time often more valuable money.
23	5.40	The poor baby as if it were sick.
31	6.50	She if she will.
35	7.40	Brothers and sisters always to help other and should quarrel.
38	8.42 weather usually a good effect one's spirits.
48	9.50	It is very annoying to tooth-ache, often comes at the most time imaginable.
54	10.76	To friends is always the it takes.

Trabue test standards. These scales are so recent that no study of their effectiveness in measuring language ability has been reported. If investigation shows, as Trabue claims, that they measure an ability that is "very closely related to what is usually called 'language ability,'" they will furnish a very convenient means of measurement. They require only a few minutes of the pupils' time and the definite instructions for scoring insure reliability. However, at best they can yield only general measures of "language ability," and some other means must be provided for diagnosis.

Trabue gives the following tentative standards for scale B, C, D, and E. These scales were constructed approximately equal in difficulty:—

<i>Grade</i>	<i>Median</i>
II.....	3.0
III.....	6.0
IV.....	8.0
V.....	9.6
VI.....	11.0
VII.....	12.3
VIII.....	13.3
IX.....	14.2
X.....	15.3
XI.....	15.8
XII.....	16.2

IV. THE MEASUREMENT OF ABILITY IN ENGLISH GRAMMAR

Types of ability. Two types of ability in English grammar may be recognized; first, the pupil's knowledge of language forms and the rules governing their use, and second, the pupil's ability to use language forms correctly. Since the function of grammatical knowledge is expected to produce approved language forms, it is the second type of ability with which the school is primarily concerned as an outcome of instruction in grammar.

Starch's Grammatical Scales. Starch¹ has devised three scales (A, B, and C) to measure a pupil's ability to use correctly certain language forms. His Grammatical Scale A consists of a series of exercises such as the following:—

¹ Starch, Daniel. "The Measurement of Achievement in English Grammar"; in *Journal of Educational Psychology*, vol. 6, pp. 615-26. Also in his *Educational Measurements*, pp. 105-08.

STEP 7

1. A fireman seldom rises above (an engineer; the position of an engineer).
2. The difference between summer and winter (is that; is) summer is warm and winter is cold.
3. He is happier than (me; I).
4. They are (allowed; not allowed) to go only on Saturday.

The pupil is given a printed copy of the scale and is directed thus: "Each of the following sentences gives in parenthesis two ways in which it may be stated. Cross out the one you think is incorrect or bad. If you think both are incorrect, cross both out. If you think both are correct, underline both." Pupils are given as much time as they need.

The sentences of the exercises have been chosen so that the difference in difficulty between any two successive steps of the scale is equal to the difference between any other two successive steps. A pupil's score is the highest step of the scale of which he does correctly three out of the four sentences. If a pupil fails on a given step, say the seventh, but does the ninth correctly, his score is 8. Thus, a pupil receives credit for each exercise of which he does correctly three out of the four sentences, but the sentences have been so arranged that only in a few cases will a pupil be able to do an exercise after he has missed the preceding.

As tentative standards of attainment Starch gives the following scores for the use of these scales:—

Grade.....	VII	VIII	IX	X	XI	XII	Freshmen
Score.....	8.0	8.3	8.6	8.9	9.2	9.5	10.3

The scale includes many different items of grammatical form and apparently these are arranged in no systematic

manner. Therefore, a pupil's score can be only a general measure of his ability to use correct language forms. To secure detailed information concerning his weaknesses it would be necessary to examine his test paper.

The Punctuation Scale. Starch has also devised a Punctuation Scale¹ which is similar in form to the Grammatical Scale. The exercises consist of sentences to be punctuated. The following extracts illustrate the nature of this scale.

STEP 6

1. We visited New York the largest city in America.
2. Everything being ready the guard blew his horn.
3. There were blue green and red flags.
4. If you come bring my book.

STEP 12

1. When thou goest forth by day my bullet shall whistle past thee when thou liest down by night my knife is at thy throat.
2. Oh come you'd better.
3. The president bowed then Hughes began to speak.

A pupil's score is determined in the same way that it was in the case of the Grammatical Scale. In the case of both scales certain features of the form of language have been isolated for the purpose of measurement. They are given in a context and in this respect the scales are similar to the sentence spelling test given on page 124. The application of these scales is simple, and the scores are reliable. The pupil's ability to distinguish certain correct forms is measured. Tentative standards of attainment are the same as for the Grammatical Scale.

The Grammar Tests. Starch has also devised three

¹ Starch, D. *Educational Measurements*, pp. 108-10.

tests for measuring directly a pupil's ability to recognize certain language forms.¹ In Test 1 the pupil is asked to mark the part of speech of each word in a certain printed text. His score is the number he designates correctly in three minutes. Test 2 calls for the designation of the case of the nouns in another printed text. Test 3 has to do with the tense and mode of verbs.

The author of these tests points out two limitations, first, "failure to cover all phases of grammatical knowledge"; and, second, "counting one designation of a part of speech, case, tense, or mode as equal to any other." The first limitation is really not a limitation of any one of the tests, but of the group of tests. The second results in considering unequal units equal which introduces a source of error.

Buckingham's Test. In making the survey of the Gary and the Prevocational Schools of New York City, Buckingham used a series of questions upon English grammar.² These questions were carefully evaluated upon the basis of difficulty. They have been rearranged and published by Haggerty.³

V. MEASURING ACCURACY IN COPYING

Copying is a phase of school work which receives little explicit attention. This probably is due to the assumption that pupils are able to copy accurately because it appears to be such a simple activity. Copying bears a relation to written expression and to other school subjects as well.

¹ Starch, D. *Educational Measurements*, pp. 110-13.

² *Seventeenth Annual Report of the City Superintendent of Schools, 1914-1915*. (Department of Education, City of New York.)

* Bureau of Coöperative Research, University of Minnesota.

Themes are usually copied before being submitted to the teacher. In solving problems in arithmetic the quantities are copied from the text. In gathering information from references copying occurs.

The Boston test. The following test of pupils' ability to copy printed matter was prepared by a group of Boston¹ teachers: —

DIRECTIONS FOR GIVING AND SCORING THE TEST

1. Read to the pupils the directions which are printed at the head of the selection they are to copy, but give them no further help. For example, do not specify possible errors which may be made.
2. Pupils ought not to see the selection until they are ready to copy it. Hence it should be placed on the desk face down until the signal is given to begin work.
3. Every error should be checked distinctly.
4. The errors which were to be noted were as follows: In spelling, capitalization, punctuation, undotted "i's," uncrossed "t's"; in omitting words, in adding words, in wrong words used, and in misplaced words.

DIRECTIONS TO PUPILS

Copy in ink as much of the following selection as you can copy accurately in fifteen minutes without hurrying. Accuracy is more important than speed: —

LIEUTENANT OULESS

In this story a young British lieutenant, in a moment of extreme irritation, strikes a private soldier. The act is one that calls for dismissal from the Queen's service. What is the officer to do? He cannot send money to the soldier — who happens to be the redoubtable Ortheris himself — nor can he apologize to him in private. Neither can he let matters drift. Ortheris, too, has his own code of pride and honor; he too is a "servant of the Queen"; but how is the insult to be atoned for? The way out of this apparently hopeless muddle is a beautifully simple one, after all. The lieutenant invites Ortheris to go shooting with him, and when they are alone, asks him "to take off his coat." "Thank you, sir!" says Ortheris. The two

¹ School Document no. 2, 1916. (Boston Public Schools, English, Determining a Standard in Accurate Copying.)

men fight until Ortheris owns that he is beaten. Then the lieutenant apologizes for the original blow, and the officer and private walk back to camp devoted friends. That fight is the moral salvation of Lieutenant Ouless. (*Bliss Perry, A Study of Prose Fiction.*)

Kinds of errors made. This test was given to 4494 first-year pupils in the Boston high schools in November, 1914, and therefore may be considered to measure the ability of pupils completing the eighth grade. The results are both interesting and significant. The following is quoted from the bulletin mentioned above:—

The errors noted consisted of nine different kinds, and the number of each kind made in this test by 4494 pupils is shown by the following tabulation:—

Spelling.....	5,829
Capitalization.....	644
Omitted words.....	4,077
Added words.....	606
Wrong words used.....	840
Misplaced words.....	105
Punctuation.....	5,876
Undotted "i's".....	8,794
Uncrossed "t's".....	606
Total.....	27,377
Average errors per pupil.....	5.54

Misspelled words. The test consisted of 170 words, 105 of them different words. It is a notable fact that every word was misspelled by somebody. It is also interesting that 92.2 per cent of the words in the test are found in Jones's *Concrete Investigation of the Material of English Spelling*.¹ In spite of the fact that these are words commonly used by children in their writing, 11.8 per cent of them were misspelled more than 100 times. This does not mean that 11.8 per cent of the children missed these words, because one pupil might have missed the same word more than once.

It is impossible to make any statement in regard to the average because many of the words occur in the selection more than once, and if misspelled by the same person each time it occurs it is counted

¹ Published by the University of South Dakota.

more than one error. Some children spelled a word incorrectly in one place and correctly in another. One boy spelled "lieutenant" wrong four out of five times, and spelled it a different way each time. Then, not all the children finished the entire selection, and no record was kept of the exact number of words each wrote. However, 4,494 pupils taking the test made 5,829 errors in spelling alone, the number of errors for each word varying from 1 to 1,045.

Undotted "i's" and uncrossed "t's." The errors made by leaving the "i's" undotted and the "t's" uncrossed comprise about one third of the entire number of errors and are largely important because of their value to legibility, as pointed out by Ayres. In connection with these errors, it is very noticeable that most of them were confined to comparatively few pupils. If a child showed a tendency to dot his "i's" and cross his "t's" in the first few lines, the chances were that that individual would have but few errors. On the other hand, if the child made many errors in the first part of the paper, there were many throughout the copying. One boy went through the entire paper without dotting an "i." Many others dotted only a small part of them.

This same test was given in Kansas City, Missouri, to the pupils in the seventh grade and in the first year of the high school. (Kansas City has only seven grades below the high school.) The average errors per pupil was 8.04 in the seventh grade, and 6.83 in the first year of high school.

VI. EDUCATIONAL SIGNIFICANCE OF THE USE OF THESE SCALES AND TESTS

Finding specific language weaknesses. The composition scales and the completion-test language scales fulfill the same function in the field of language as the handwriting scales of Ayres and Thorndike do in that field. They are instruments for general measurement. By means of them a teacher can obtain a measure of the language abilities of her pupils in terms of fixed units which she may compare with established standards or with similar measures of other

groups of pupils. They also indicate those pupils who are below standard and who for this reason need instruction. However, before this instruction can be intelligently given the language ability of these pupils must be diagnosed to locate the exact defects.

The grammar tests and the copying test measure specific abilities. They furnish the teacher with detailed information about the several members of her class. If a pupil is weak in punctuation that fact is revealed. The teacher then knows that she must instruct that pupil in punctuation. If a pupil makes certain types of errors in copying, he needs certain instruction.

Remedying the situation revealed. When a teacher learns the specific language weaknesses of her pupils she is then in position to apply more intelligently her stock of methods and devices of instruction. In language as in the case of the other subjects, the teacher must instruct individual pupils who are grouped together rather than groups of pupils. Furthermore, each pupil should receive the instruction which he needs to correct his language errors.

If pupils are weak in a language ability, such as punctuation, the laws of habit-formation apply. After being sure that he understands the function of the punctuation marks, a pupil must have practice in punctuating his own writing. This probably is not sufficient. Exercises for practice can be constructed by taking appropriate material and reproducing it without the punctuation marks.

Until a teacher recognizes definite and specific ends to be attained there is certain to be a large degree of dissipation of her efforts. Perhaps one reason why language in-

struction so often does not produce satisfactory results is that it is not directed toward the engendering of definite abilities. That our present standards of language are chaotic is indicated in the report of a recent investigation.¹

Analyzing language ability. The six compositions comprising the Harvard-Newton Exposition Scale were reproduced without any identifying marks. They were graded on the scale of 100 per cent by twenty-four eighth-grade teachers who were asked to follow certain typewritten directions. The six compositions were then "completely corrected so far as mechanical or measurable errors were concerned." The corrected compositions were graded by the same teachers according to the same directions.

If the "mechanical errors" of the compositions were significant factors in determining the first set of marks, the second set of marks should be conspicuously higher. However, this was not the case. For two of the compositions the average "grade" was less after the "mechanical errors" had been corrected. The individual marks show that some teachers consider form important, and that others tend to disregard it in marking a composition.

In teaching spelling teachers have kept a record of pupils' errors and have emphasized these words in their teaching. In our consideration of spelling it was urged that teachers first ascertain what words their pupils were unable to spell correctly. This plan may be adapted to the teaching of other aspects of language. The teacher should ascertain the pupils' grammatical errors, and then equip them with

¹ Brownell, Baker. "A Test of the Ballou Scale of English Composition"; in *School and Society*, vol. 4, pp. 938-42.

the rules of grammar which are needed to correct them. This has been done on a large scale in St. Louis and Kansas City, Missouri.¹

Perhaps the scales and tests described in this chapter will have fulfilled their most important function if they cause teachers to analyze and define "language ability" in more specific terms. It is believed that their use will tend to produce this result. Analysis of "language ability" and specific definition of the elements are greatly needed. Upon the accomplishment of these two things depends the construction of more valuable measuring instruments in the language field and the scientific determination of methods and devices of instruction.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How does the problem of measurement in language differ from the problem of measurement in arithmetic?
2. What makes the problem of measurement in language difficult?
3. How does the Hillegas Scale differ from the Willing Scale? The compositions which compose the Willing Scale were written under defined conditions and on similar topics. Does this make it a superior scale?
4. Which of the composition scales described in this chapter will be the most helpful to the teacher? Why?
5. Give the copying test to your pupils following the directions carefully. Do the results agree with your estimate of the ability of your pupils to copy?
6. Keep accurate lists of the language errors of your pupils. What are the rules which are necessary to correct these errors? Are they the rules upon which you are placing the most emphasis in your teaching?
7. Do you have definite objective standards of attainment in English composition? Can you use the tests described in this chapter to establish such standards?
8. Do you think pupils would be helped by having definite objective standards of attainment established for them?

¹ See report by W. W. Charters in the *Sixteenth Yearbook of the National Society for the Study of Education*, part I.

BIBLIOGRAPHY

Only the most important references are given here. Additional references will be found in the footnotes in the chapter.

I. GRAMMAR

1. *Starch's Grammatical Scale A.* Copies may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.
2. *Starch's Punctuation Scale A.* Copies may be obtained from the above address.
3. *Starch's English Grammar Test 1, 2, and 3.* Copies may be obtained from the above address.

REFERENCE: Starch, Daniel. "The Measurement of Achievement in English Grammar"; in *Journal of Educational Psychology*, vol. 6, pp. 615-25. See also Starch, Daniel, *Educational Measurements*.

4. *Haggerty's English Grammar Tests.* An arrangement of Buckingham's English Grammar Tests. Copies may be purchased from the Bureau of Coöperative Research, University of Minnesota, Minneapolis, Minn.
5. *Buckingham's English Grammar Tests.* Used in the Survey of the Gary and Prevocational Schools of New York City.

REFERENCE: *Seventeenth Annual Report of the City Superintendent of Schools*, New York City, 1914-15.

II. ENGLISH COMPOSITION

1. *Harvard-Newton Composition Scale*, devised by F. W. Ballou. Copies of the scale may be secured from the Harvard University Press, Cambridge, Massachusetts.

REFERENCES: Hudelson, Earl. "Some Achievements in the Establishment of a Standard for the Measurement of English Composition in the Bloomington, Indiana, Schools"; in *English Journal*, vol. 5, pp. 590-97. (November, 1916.)

Kayfetz, Isidore. "A Critical Study of the Harvard-Newton Composition Scales"; in *Pedagogical Seminary*, vol. 23, pp. 325-47. (September, 1916.)

2. *Hillegas's Composition Scale.* Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCES: "Hillegas Scale for Measurement of English Composition"; in *Teachers College Record*, September, 1912.

Johnson, F. W. "The Hillegas-Thorndike Scale for Measuring the Quality in English Composition by Young People"; in *School Review*, vol. 21, pp. 39-49.

Kayfetz, Isidore. "A Critical Study of the Hillegas Composition Scale"; in *Pedagogical Seminary*, vol. 21, pp. 559-77.

Thorndike, E. L. "Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition"; in *English Journal*, vol. 2, p. 551.

3. *Thorndike Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People*. No directions for use. Copies may be obtained from Bureau of Publications, Teachers College, Columbia University, New York City.

4. *A Scale for Measuring the General Merit of English Composition in the Sixth Grade*, by F. S. Breed and F. W. Frostic. See *Elementary School Journal*, vol. 17, pp. 307-25.

5. *M. H. Willing's Composition Scale*.

REFERENCES: *Report of the School Survey of School District Number One in the City and County of Denver*, part II, p. 59.

Willing, M. H. *Measurement of Written English Composition in the Public Elementary Schools of Denver, Colorado*. (Master's Thesis, University of Chicago. Unpublished.)

6. *The Nassau County Supplement to the Hillegas Scale*, by M. R. Trabue. Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCE: Trabue, M. R. "Supplementing the Hillegas Scale"; in *Teachers College Record*, vol. 18, p. 51. (January, 1917.)

7. *Trabue's Completion-Test Language Scales*. Copies may be obtained from the Bureau of Publications, Teachers College, Columbia University, New York City.

REFERENCE: Trabue, M. R. *Completion-Test Language Scales*. (Teachers College Contributions to Education, no. 77, 1916.)

CHAPTER VII

HIGH-SCHOOL SUBJECTS

THE preceding five chapters have dealt with the "tool" subjects in the field of elementary education. In each case the outcomes of instruction were specific habits. In the high-school field other outcomes of instruction predominate. These outcomes of instruction are less definite and tangible, and consequently the problem of measurement involves greater difficulties. This condition, coupled with certain others, probably accounts for the scarcity of tests in the high-school field. In this chapter a brief description is given of the tests which have been announced. In only a few cases have the tests been used sufficiently to give an indication of their effectiveness.

I. ALGEBRA

The problem of measurement. As in treating other subjects, the first step in considering the problem of measurement in the case of algebra is to determine the fundamental outcomes to be realized in the teaching of the subject. Among the specific outcomes¹ are the comprehension and manipulation of symbols in performing the operations of elementary algebra. These operations are used as tools in the solution of equations and of "practical" problems. The manipulation of symbols must be automatic for the

¹ For a complete statement of the outcomes of instruction in elementary algebra see Rugg, H. O. "The Experimental Determination of Standards in First-Year Algebra"; in *School Review*, vol. 25, pp. 37-66; 196-213.

same reason that the most commonly used words must be spelled automatically or that the operations of arithmetic must be performed without focusing the attention upon the activity. As in the case of arithmetic each type of example requires a specific ability. Hence it is necessary to determine the fundamental types of examples in elementary algebra.

The fundamental operations of algebra. The plan of our texts suggests that algebra is simply generalized arithmetic, that is, that it consists primarily of performing the operations of addition, subtraction, multiplication, and division with algebraic symbols instead of with the Hindu characters used in arithmetic. Thus the fundamental operations of algebra appear to be addition, subtraction, multiplication, and division. This is not entirely true. The equation is the principal tool or instrument of algebra in dealing with "practical" problems. Practically all of elementary algebra is grouped about the equation. In fact algebra has been defined as the science of the equation. From this point of view the fundamental operations of algebra are those required in the solution of equations.

In elementary algebra a very large per cent of the "practical" problems require only simple equations. Many of the simple equations will contain fractions with numerical denominators, but very few of the problems will result in fractional equations having the unknown quantity in the denominator.¹ Therefore, from the standpoint of usefulness

¹ See Monroe, Walter S. "An Experiment in the Organization and Teaching of First-Year Algebra"; in *School Science and Mathematics*, vol. 12, pp. 225-31.

the first group of fundamental operations of elementary algebra are those which occur in the solution of simple equations containing fractions with numerical denominators. The operations which are necessary for solving quadratic equations and simultaneous equations would form other groups of fundamental operations. To learn the group of fundamental operations which are required to solve simple equations consider the solution of this equation:

$$-\frac{6x - 1}{8} - (-4x - 8) = \frac{5(3x + 4)}{6}$$

Clearing equation of fractions

$$-18x + 3 + 96x + 192 = 60x + 80$$

Transposing terms

$$-18x + 96x - 60x = 80 - 3 - 192$$

Collecting Terms

$$18x = -115$$

Finding value of x . $x = -\frac{115}{8}$

The operations involved in the solution of this equation are: (1) clearing equation of fractions, (2) transposing terms, (3) collecting terms, and (4) finding value of x . Clearing an equation of this type of fractions involves the multiplication of a binomial by an integer, and under certain conditions the multiplication of a binomial by a binomial. Collecting terms is a very simple type of addition and subtraction. Finding the value of x is a special form of division. In the solution of some non-fractional equations the multiplication of one binomial by another will occur.

In a similar manner the operations which occur in the solution of quadratic equations and simultaneous equations may be determined. The operations required in the solu-

tion of these two types of equations, together with those required for the simple equation, constitute the fundamental operations of elementary algebra. There will be no need for factoring until quadratic equations are taken up, and the operation is not necessary then. Exponents beyond the square are rarely used.

At least one attempt¹ has been made to determine the fundamental operations of algebra by analyzing the content of currently used textbooks. The fundamental operations determined by this method are given as follows:—

- | | |
|---------------------------|--|
| 1. Removal of parentheses | 8. Clearing of fractions and
fractional equations |
| 2. Combining terms | 9. Quadratic equations |
| 3. Subtraction | 10. Graphing of equations |
| 4. Evaluation | 11. Solution of "practical" for-
mulæ |
| 5. Special products | 12. Simultaneous equations |
| 6. Factoring | |
| 7. Exponents | |

This determination is subject to the same limitations as must be placed upon all statements of the consensus of present practice. It involves the *a priori* acceptance of the traditional subject-matter of elementary algebra as a satisfactory basis.

In order that the results of using a test may be significant to the teacher, it is necessary that it measure the ability to do something which has a fundamental importance. For example, a test which measures the ability of pupils to perform long division in algebra furnishes the teacher with information of little importance, because long division is not

¹ Rugg, H. O., and Clark, J. R. "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 137-66; 196-213.

used in dealing with problems. In fact an examination of an algebra text will reveal that the pupil has practically no opportunity to use this operation after he leaves the topic of long division.

Standard Research Tests in Algebra. Upon the basis of the analysis of the solution of simple equations the Standard Research Tests in Algebra¹ were constructed. These tests consist of a series of six tests. Each of the first five tests is designed to measure the ability to do one of the operations occurring in the solution of simple equations.

The tests are:—

Test I. $\pm a (\pm bx \pm c)$, a , b , and c , being not greater than 9 and not all positive.

Test II. Clearing equations of fractions.

Test III. Solving for x , a special case of division.

Test IV. Transposition.

Test V. Collecting terms, a special case of addition and subtraction.

Test VI. Simple equations to be solved.

In giving the tests each pupil is provided with a printed copy of the exercises to be done. A definite time is allowed for each test. The ability of a pupil is measured by the number of exercises he does in a given time, and by the accuracy of his work.

Other algebra tests. A number of other tests have been devised to measure algebraical abilities.

Thorndike has devised a test which consists of a series of exercises arranged in order of increasing difficulty. The relative difficulty was determined on the basis of the judg-

¹ Monroe, Walter S. "A Test of the Attainment of First-Year High-School Students in Algebra"; in *School Review*, vol. 23, pp. 159-71.

ments of two hundred teachers of algebra. The scale is as follows:¹

If $a = 4$ and $b = 2$, what does $a + b$ equal. Answer.....

If $a = 4$ and $b = 0$, what does $a + b$ equal? Answer.....

If $x + 3a = 5a$, what does x equal? Answer.....

Find the average midnight temperature for the week in which the daily midnight temperatures were 15, 3, 0, -7, -9, 6, and 17 degrees. Answer.....

If $\frac{1}{a} - \frac{1}{x} = \frac{1}{x} - \frac{1}{b}$, what does x equal? Answer

$$\frac{x}{a} - 1$$

If $2 + \frac{2}{\frac{a}{a}} = 0$, what does x equal? Answer

A man has a hours to spend riding with a friend. How far can they ride together, going out at the rate of b miles an hour and just covering the return trip at the rate of c miles an hour. Answer.....

How much water must be added to a pint of "alcohol, 95 per cent pure" to make a solution of alcohol, "40 per cent pure"? Answer.....

Given that $2x - 3$ is less than $x + 5$, and that $11 + 2x$ is less than $3x + 5$, to find the limits (i.e., the values) between which x lies.

Two series of algebra tests, consisting of 12 tests each have been devised. One has been devised by *Rugg and Clark*² and the other by *Childs*.

In this second series the Standard Research Tests described above were incorporated. It is interesting to note that the two series of tests, each consisting of the same number of separate tests, agree on only five operations: (1) multipli-

¹ So far as the writer knows copies of the test are not available for distribution. See Rugg, H. O. "The Experimental Determination of Standards in First-Year Algebra"; in *School Review*, vol. 25, p. 43.

² *Loc. cit.*, pp. 37-66. In the most recent publication of these tests the number of tests in this series has been increased to sixteen.

cation of a binomial by an integer, (2) factoring, (3) solution of simple equations, (4) stating problems, (5) solving simultaneous equations. The series devised by Rugg and Clark include tests for the following additional operations: (1) product of two binomials, (2) evaluation of algebraic expressions by substitutions, (3) involution and operations based on other laws of exponents, (4) quadratic equations, (5) graphical solution of simultaneous equations, (6) evolution, (7) solving formulæ for certain quantities. The series devised by Childs include tests for the following operation in addition to those which are common to the two series: (1) addition and subtraction, (2) finding the value of x in such equations as $5x = -31$, (3) long division, (4) transposition, (5) collecting terms, (6) clearing equations of fractions. (Two tests are devoted to the stating of problems.)

Conclusions from the tests. The fact that there are so few points of agreement among those who have devised tests in the field of algebra is significant. It indicates, in the first place, that the fundamental operations of algebra are not simply addition, subtraction, multiplication, and division, as in arithmetic. In the second place it indicates that the textbook writers and teachers of algebra have not yet determined the operations which are fundamental. In the analysis of the subject-matter of algebra given above the writer has indicated his belief that the fundamental operations of algebra are those required in the solution of equations. The acceptance of this premise furnishes a definite basis for constructing a series of tests.

Standards. The Standard Research Tests in Algebra have been given in a number of high schools. The following

median scores, based upon returns from twenty-one cities, may be taken as tentative standards:—

<i>Test</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>
Number of Pupils	2077	1993	2107	2127	2198	1992
Speed, number of examples attempted.	14.6	5.4	11.5	10.2	11.2	8.3
Accuracy, per cent of examples correct	96	41	100	94	77	32

Tentative standards for certain tests of the series devised by Rugg and Clark are given in the *School Review*, vol. 25, pp. 122-23. These standards are also printed on the class record sheet which is furnished with the tests. Childs has recently published a report of the use of his series of algebra tests in fifteen cities.¹

Meeting the teaching situation revealed by algebra tests. The mental processes of algebra are similar to those of arithmetic in many respects. In both subjects the operations must be performed automatically in order to free the attention for the doing of other things. The abilities are specific. In arithmetic it was shown that addition involved not simply one ability but several abilities. (See page 18.) Each type of example required a specific ability. A scientific analysis of algebraical abilities has not been made, but it is probable that in algebra each type of example requires a specific ability if it is done automatically. The engendering of arithmetical abilities is based upon the laws of habit formation. These laws also apply to the teaching of the

¹ Childs, Hubert G. "The Measurement of Achievement in Algebra," in the Third Conference on Educational Measurements. (*Bulletin, Extension Division, Indiana University*, vol. 2, no. 6, pp. 171-83.)

operations of algebra. Individual differences complicate the teaching of arithmetic. They have been shown to be equally conspicuous in algebra.

Class instruction possesses the same weaknesses in algebra as it does in arithmetic. The writer has observed the plan of giving drill, described on page 55, in algebra classes as well as in arithmetic. The result was the same in both. Each pupil needs drill upon the types of examples he does not do well. Practice tests for algebra¹ can be devised upon the same principles as those for arithmetic. The suggestions given on page 59, for adapting the instruction to the needs of the pupils, may be applied to instruction in algebra as well.

Rugg and Clark state that "It has been shown that success in teaching algebra depends primarily on the teacher's knowledge of the typical difficulties which the pupils will meet in learning algebra." In the reports of the use of algebra tests the errors made by pupils have been studied. In all cases the number of errors has been large. Table XXV gives the types of errors which were made by two hundred and seventy-five first-year pupils on the Standard Research Tests in Algebra. The tests were given in March, 1914. Rugg and Clark have used a more elaborate classification of errors, but their results indicate the same condition.

The conditions revealed clearly indicate that the pupils have not been given sufficient satisfactory drill to make automatic the performing of the simpler operations of algebra. Before this can be accomplished, teachers must determine what the important or fundamental operations of algebra

¹ See Rugg, H. O., and Clark, J. R. "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 196-213. (March, 1917.)

are. When this has been accomplished they must further determine what types of exercises occur in each operation. For example, assuming that performing the indicated operation of $a(bx + c)$ is fundamental, it is obvious that these types of exercises occur, $a(bx + c)$, $-a(bx + c)$, $a(-bx + c)$, $-a(-bx + c)$, $a(bx - c)$, $-a(bx - c)$, $-a(-bx - c)$, $a(-bx - c)$. In the studies referred to it was found that each required its own specific ability. This being the case the teacher must provide satisfactory drill upon each type. A series of scientifically constructed practice exercises furnish a means for doing this.

TABLE XXV. CLASSIFICATION OF ERRORS MADE BY TWO HUNDRED AND SEVENTY-FIVE FIRST-YEAR PUPILS ON THE STANDARD RESEARCH TESTS (ALGEBRA)

<i>Test</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>
Mistake in sign.....	436	265	184	295	387	739
Mistake in the common denominator or in its use.....		167				371
Mistakes in arithmetic.....	143	249	498		596	391
Mistakes in copying.....				382	63	82
One term of binomial not multiplied.....	29					16
A term neglected.....		103		302	86	26
x omitted.....	117	19		53	80	16
Incomplete as $-x = 5$			38			26

II. GEOMETRY

A geometry test of seventy questions, arranged in twenty groups, has been devised by Stockard and Bell.¹ "These

¹ Stockard, L. V., and Bell, J. C. "A Preliminary Study of the Measurement of Abilities in Geometry"; in *Journal of Educational Psychology*, vol. 7, pp. 567-80.

groups involve drawing figures, naming figures, indicating order of development in demonstration, completing statements, stating of converse, definitions, regular polygons, parts of a demonstration, angular relations, area of trapezoid, angles in polygons, angles in circles, congruency of triangles, similarity of triangles, loci, auxiliary lines, simple constructions, ratio and proportion, algebraic expression of geometrical relations, and equivalent construction." The questions are asked in such a way that many pupils are able to complete the list in forty minutes. On the basis of the results of giving the tests to 372 students the difficulty of the questions has been expressed in terms of a common unit.

III. FOREIGN LANGUAGES

Foreign languages involve some of the same mental processes as occur in the field of the English language. A pupil must associate the correct meaning with words and groups of words. In using a foreign language to express ideas, the rules of grammar which govern the form must be followed. Hence the problem of measurement is similar in several respects.

Starch's foreign language tests. Starch has devised both vocabulary and reading tests for Latin, German, and French.¹ The same plan has been followed for each language. Only the Latin test will be described here. The words of the Latin vocabulary test were selected from Harper's Latin Dictionary by taking the first word on every twentieth page. This gave a list of one hundred words. In addition to the

¹ Starch, D. *Educational Measurements*, pp. 171-87.

Latin words the English equivalents are given on the test paper. Both lists are arranged alphabetically. The test consists of associating with each Latin word its English equivalent. The Latin reading test consists of sentences selected from first-year Latin books, Cæsar, Cicero, and Virgil, and the reading tests in German and French are composed of simple sentences. The sentences in each of the tests are arranged in order of increasing difficulty.

The vocabulary test measures the extent of the pupil's "vocabulary." However, it may be that this "vocabulary" is not the same as the vocabulary which he is able to use in translating Latin into English. Both the vocabulary tests and the reading tests are simple to use.

The Hanus Latin tests. Hanus has also devised a group of Latin tests¹ which consists of four tests for vocabulary, a translation test, and a grammar test. All of these tests are based on Cæsar and Cicero. No words appear in the vocabulary tests "which occur less than one hundred times in Cæsar and Cicero." The translation test "contains only constructions which are found at least five hundred times in Cæsar and Cicero." The grammar test is based on the sentences to be translated. The vocabulary test differs from the one devised by Starch in that the pupil must write the English equivalent.

Henmon's Latin tests. V. A. C. Henmon, of the University of Wisconsin, has devised a series of Standard Tests in Latin which consist of three tests. *First*, an easy hundred-word vocabulary test — fifty in English and fifty in Latin —

¹ Hanus, Paul. "Measuring Progress in Learning Latin"; in *School Review*, vol. 24, pp. 342-51,

containing the words that are common to four widely used first-year books. *Second*, a Standard Vocabulary Test of 239 words representing all the words common to thirteen first-year books and to Cæsar, Cicero, and Virgil. *Third*, the Latin sentence test consists of thirty sentences constructed by using none but the 239 words of the Standard Vocabulary Test.

A pupil's vocabulary is important in the study of a foreign language, as well as in the field of English. Hence in using a vocabulary test the teacher is measuring an important ability. The test will reveal that some pupils have much smaller vocabularies than others. By comparing the class score of his pupils with the standard a teacher may know whether he is placing sufficient emphasis upon the feature of the language which the test measures. A translation test gives a general measure of the ability of a pupil to translate. It also furnishes the teacher with a means for comparing her class with an objective standard.

No devices for remedying the shortcomings which tests reveal have been worked out. Several of the devices given in the treatment of reading and language can easily be adapted to the teaching of a foreign language. As in the case of other subjects, the tests will fulfill an important function if they do nothing more than demonstrate to the teacher that he is instructing a group of pupils who differ widely, and that his efforts will be most effective when each pupil is given the instruction which he needs.

IV. PHYSICS

Starch has also devised a series of tests in physics, covering mechanics, heat, sound, light, and magnetism and

electricity.¹ The tests consist of sentences, from which words have been omitted. The sentences and the words to be omitted have been chosen so that a pupil cannot supply the correct words unless he knows certain physical facts or principles. The following sentences will illustrate the tests:—

17. The periods of pendulums of equal lengths swinging through short arcs are independent of and also independent of
30. The point degrees below degrees centigrade is called
39. The frequency of vibration of a string varies inversely as
53. The critical angle is that angle of incidence which will produce
55. Electromotive force is the difference in between

The tests consist of seventy-five mutilated sentences of this type. The facts, principles, and laws upon which these sentences are based were determined by examining five widely used textbooks. The one hundred and two facts, principles, or laws which were treated by all five of the textbooks are the ones which the pupils must know to do the tests correctly. These facts and principles probably are the ones fundamental to elementary physics. If the tests do nothing more than define the fundamental facts and principles they will fulfill an important function. However, they also give the teacher a means for comparing his class with other classes.

¹ Starch, D. *Educational Measurements*, pp. 188-92.

V OTHER TESTS WHICH MAY BE USED IN THE HIGH SCHOOL

Certain tests of those described in the preceding chapters are intended to be used in the high school, as well as in the elementary school. These are Test III, of the Kansas Silent Reading Tests; the Thorndike Scale Alpha for Measuring the Understanding of Sentences; Starch's English Vocabulary Test; the composition scales; the copying test; the Trabue Completion-Test Language Scales; and Starch's Grammatical Tests. For the description of these tests and the use of them the reader is referred to the preceding chapters.

In addition to these tests many of the others have been applied to high school pupils. For example, the Courtis Standard Research Tests in Arithmetic, Series B, have frequently been given to high school pupils, although many of them were not studying arithmetic. However, in applying such tests to high-school pupils it should be remembered that the tests were not designed for that purpose. Consequently it may be expected that the tests will not be as satisfactory as when used in the way they were intended to be used.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. How do you account for the fact that less progress has been made in devising tests for high-school subjects than for elementary school subjects?
2. Which of the algebra tests described in this chapter do you think would be most helpful to a teacher? Why?
3. Compare the Latin vocabulary tests which are described. Which one do you think would be most helpful to a teacher? Why?

4. Is a test like the physics tests superior to an ordinary examination? Justify your answer.
5. The geometry test described covers a very wide range of topics. Is this necessary for a test in such subjects as geometry, physics, or history?
6. If you are teaching algebra, study the errors which your pupils make. Suggest a plan for reducing the number of errors.
7. Suggest several plans for increasing the foreign language vocabulary of pupils who have been shown to be below standard. How could you determine which of the plans is the best.

BIBLIOGRAPHY

I. ALGEBRA

Only the most important references are given here. Additional references will be found in the footnotes of the chapter.

1. *Indiana Algebra Tests*, arranged by H. G. Childs. Copies may be obtained from the University Book Store, Bloomington, Indiana.

REFERENCE: Childs, Hubert G. "The Measurement of Achievement in Algebra"; in *Third Conference on Educational Measurements*. (Bulletin of the Extension Division, Indiana University, vol. 2, no. 6, pp. 171-83.)

2. *Preliminary Algebra Tests*, devised by C. E. Stromquist. Copies may be obtained from the University of Wyoming, Laramie, Wyoming.

3. *Standardized Tests in First Year Algebra*, devised by H. O. Rugg and J. R. Clark. Copies may be obtained from H. O. Rugg, School of Education, University of Chicago.

REFERENCES: Rugg, H. O. "The Experimental Determination of Standards in First Year Algebra"; in *School Review*, vol. 24, pp. 37-66. (January, 1916.)

Rugg, H. O., and Clark, J. R., "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 113-32 and 196-213. (February and March, 1917.)

4. *Standard Research Tests in Algebra*, devised by Walter S. Monroe. Copies may be obtained from the Bureau of Educational Measurements and Standards, Emporia, Kansas.

REFERENCE: Monroe, Walter S. "A Test of the Attainment of First-Year High-School Students in Algebra"; in *School Review*, March, 1915.

5. *A Scale for Testing Ability in Algebra*, devised by W. H. Coleman. Copies may be obtained from W. H. Coleman, Bertrand, Nebraska.
6. *Thorndike's Algebra Test*.

REFERENCE: Rugg, H. O. "The Experimental Determination of Standards in First-Year Algebra"; in *School Review*, vol. 25, pp. 37-66. (January, 1916.)

II. GEOMETRY

REFERENCE: Stockard, L. V., and Bell, J. C. "A Preliminary Study of the Measurement of Abilities in Geometry"; in *Journal of Educational Psychology*, vol. 7, pp. 567-80.

III. FOREIGN LANGUAGES

1. *French Vocabulary and Reading Tests*, devised by Daniel Starch.
2. *German Vocabulary and Reading Tests*, devised by Daniel Starch.
3. *Latin Vocabulary and Reading Tests*, devised by Daniel Starch.

Copies of all three of these tests may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.

REFERENCE: Starch, Daniel. *Educational Measurements*, chaps. xi, xii, and xiii. (The Macmillan Company.)

4. *Hanus's Latin Tests*.

REFERENCE: Hanus, Paul H. "Measuring Progress in Learning Latin"; in *School Review*, vol. 24, pp. 342-51. (May, 1916.)

5. *Henmon's Latin Tests*. Copies of the tests may be obtained from V. A. C. Henmon, University of Wisconsin, Madison, Wisconsin.

IV. PHYSICS

1. *Tests for Physics*, devised by Daniel Starch. Copies may be obtained from Daniel Starch, University of Wisconsin, Madison, Wisconsin.

REFERENCE: Starch, Daniel. *Educational Measurements*, chap. xiv. (The Macmillan Company.)

CHAPTER VIII

STATISTICAL METHODS

THE teacher, principal or school officer making use of the standard tests previously described has but little use for any mathematical knowledge of statistical methods. The little knowledge that is needed relates much more to proper methods of tabulating and the most effective methods of graphic representation. The few calculations that need to be made are arithmetical, and the statistical terminology needed is quite small. This chapter is intended to state briefly what will be needed, and to give a few typical illustrations of good methods in tabulating and charting.

Good arrangement of scores. The significance of a group of facts, such as the scores made by a class upon a test, may be made more evident by certain methods of arranging them, and by determining an index of the central tendency of the group. Take, for example, the scores which were made by a sixth-grade class of thirty-five pupils when given the Kansas Silent Reading Test. When these scores are presented in the manner of Table XXVI the array tends to confuse. One must scan the entire array to learn that the lowest score is 4.2, or that the highest score is 30.1. One cannot easily learn that pupil BB, who made a score of 14.9, stands eighth from the poorest in the group. If now the scores are simply rearranged in order of magnitude, as shown in Table XXVII, their significance is much more easily grasped.

TABLE XXVI. SHOWING A POOR ARRANGEMENT OF SCORES

Pupil	Score	Pupil	Score	Pupil	Score
A	27.3	M	10.0	Y	16.0
B	19.2	N	16.3	Z	19.1
C	26.2	O	21.1	AA	15.4
D	22.5	P	25.6	BB	14.9
E	15.4	Q	21.1	CC	16.4
F	18.3	R	15.9	DD	14.1
G	28.4	S	16.1	EE	4.2
H	17.4	T	5.9	FF	20.0
I	25.1	U	30.1	GG	24.1
J	15.7	V	22.3	HH	26.3
K	11.8	W	13.1	II	25.8
L	21.6	X	12.8		

TABLE XXVII. SHOWING THE SAME SCORES REARRANGED IN A BETTER ORDER

Pupil	Score	Pupil	Score	Pupil	Score
EE	4.2	Y	16.0	V	22.3
T	5.9	S	16.1	D	22.5
M	10.0	N	16.3	GG	24.1
K	11.8	CC	16.4	I	25.1
X	12.8	H	17.4	P	25.6
W	13.1	F	18.3	II	25.8
DD	14.1	Z	19.1	C	26.2
BB	14.9	B	19.2	HH	26.3
AA	15.4	FF	20.0	A	27.3
E	15.4	O	21.1	G	28.4
J	15.7	Q	21.1	U	30.1
R	15.9	L	21.6		

The median. The eighteenth score is the middle one of this group. Such a score is called the "median score," and indicates the "central tendency" of the group. The general standing of a group of pupils is expressed by the central tendency. In case there is an even number of scores, there is no middle score. In such a case the average of the two most central scores may be taken, although for practical

purposes it will be satisfactory to take the lesser of the two middle scores. Thus, if a distribution contains forty-one scores, the middle score is the twenty-first score; if it contains forty scores, the twentieth score may be taken as the middle score.¹ The number of the middle score is obtained by dividing the total number of scores by two. This quotient expressed as the nearest integer is called the "half sum." When using a particular test the directions which accompany that test should be followed, because the medians used for standards have been obtained by following the accompanying directions.

Frequency of scores. Usually scores are expressed simply in whole numbers. When this is the case, the distribution is simply the statement of how many scores there are of each magnitude. This is done as in Table XXVIII. The column labeled "Score" gives the scores arranged in order of magnitude. The column headed "Frequency" tells how many scores there are of each magnitude, or how frequently each score occurs. In this table there are two scores of 8, one score of 7, four scores of 6, etc. The total number of scores is 33.

¹ There is a difference of practice on this point. The directions which accompany the Kansas Silent Reading Tests read as follows:—

"The median score is the score on the middle paper in the pile of papers arranged according to size of scores. If there are thirty-five papers, the median score is the score on the eighteenth paper. If there are thirty-six papers, the median score is half way between the score on the eighteenth paper and the score on the nineteenth paper."

The directions which accompany the Courtis Standard Research Tests in Arithmetic, Series B, read as follows:—

"If there are thirty-seven children in the class, the nineteenth score in order of magnitude would be the median score; for there would be eighteen scores larger and eighteen smaller. If there were thirty-six children in the class, the eighteenth score would be taken as representing the nearest approximation to the middle measure."

TABLE XXVIII. SHOWING THE DISTRIBUTION OF SCORES

<i>Score</i>	<i>Frequency</i> ¹
12	—
11	—
10	1
9	—
8	2
7	1
6	4
5	6
4	8
3	6
2	4
1	1
0	—
Total	33

Here the seventeenth score is the middle one, but it is not possible to identify its value directly. It is clearly one of the eight scores given as 4, because counting from the lower end of the distribution, 1 and 4 are 5, and 6 are 11, and 8 are 19, which is beyond the middle of the distribution. The approximate value of this median, that is, the seventeenth score is 4.0. To determine the approximate median it is simply necessary to locate the interval of the distribution in which the median falls.

To locate the interval in which the median falls, begin at the lower end of the distribution and add together the frequencies until the addition of the next one will make a sum greater than the number of the median score, or half of the total of the frequencies. This sum of the frequencies is called the partial sum. The median score is in the next interval, and the approximate median is the value of that interval.

Intervals of distribution. In Table XXVII each score is

¹ Frequency or number of pupils making the score.

listed separately. For certain purposes this makes the distribution more complex and awkward to work with than if the scores were classified in a few groups. A convenient choice of groups for this array of scores would be as shown in Table XXIX. In the group or interval from 14.0 to 15.9 there are six scores — 14.1, 14.9, 15.4, 15.4, 15.7, and 15.9. When they are grouped together each score loses its identity, but it should be remembered that the scores which are grouped together are not necessarily equal. In calculating the true median of a distribution, it is necessary to assign an identity to the middle score. We do this by assuming that the exact values of the scores included within an interval are uniformly distributed over the interval.

TABLE XXIX. ARRANGEMENT OF SCORES BY INTERVAL GROUPS

<i>Intervals</i>	<i>Frequency</i>
30.0 to 31.9	1
28.0 to 29.9	1
26.0 to 27.9	3
24.0 to 25.9	4
22.0 to 23.9	2
20.0 to 21.9	4
18.0 to 19.9	3
16.0 to 17.9	5
14.0 to 15.9	6
12.0 to 13.9	2
10.0 to 11.9	2
8.0 to 9.9	
6.0 to 7.9	
4.0 to 5.9	2
2.0 to 3.9	
.0 to 1.9	
Total	35

Many of the tests yield scores in terms of the intervals of distribution. This is true of the scores from the Courtis

Standard Research Tests, Series B. The pupil's score is so many examples attempted and so many right. From this it may appear that the scores should not be considered to be distributed over the interval, but it must be remembered that in marking the test papers no credit was given for examples partly completed or for examples partly right. Thus all fractions have been dropped. This being true, it is obvious that the accurate measures of the pupils are really probably distributed uniformly over the interval.

Approximate and true median. To calculate the amount to be added to the approximate median to make the true median, proceed as follows: (1) Subtract the partial sum of the frequencies from the half sum. The partial sum is found in determining the approximate median. (2) Divide this difference by the number of scores which are included in the interval in which the true median falls. Add this quotient to the approximate median. If the width of the interval is more than one unit, as in Table XXIX, the quotient must be multiplied by the number of units the interval contains. In the case of Table XXIX the quotient would be multiplied by 2. It is well to carry the quotient to two decimal places, but in writing the median it should be expressed only to the nearest tenth.¹

The approximate median of the distribution given in Table XXIX is 18.0. The half sum is 18 and the partial sum is 17. The difference is 1, and this, divided by 3, the number of scores in the next interval, give a quotient of .33. Since the width of the interval is 2, .33 is multiplied

¹ See King, W. I. *The Elements of Statistical Method*, pp. 129-30; and Thorndike, E. L., *Mental and Social Measurements*, p. 54.

by 2. A correction of .66 is to be added to the approximate median, 18.0. This gives 18.66, which should be written as 18.7, the true median of the distribution.

The average. Another central tendency is the average or arithmetical mean. This is, however, much less used than the median, and usually is a much less accurate measure to use. The average of a group of scores may be found by dividing the sum of the scores by the number of scores. However, if the group is a large one, this involves considerable labor. In such a case a short method may be used.¹

The mode. Still another central tendency, though not much used, is the mode. By mode we mean the most common score or measure. Thus, in Table XXVIII the mode is 4.0 and in Table XXIX it is 14.0. Average and mode, though, are but little used, the most commonly used central tendency measure being the median.

Measures of variability; (1) average deviation. Both the median and the average represent the central tendency of the magnitude of the group of scores. Frequently it is helpful to obtain an index of the variability of the scores, that is, of the closeness with which the scores are grouped about the median or the average. The amount by which a score differs from the central tendency is called the deviation. A score which is greater than the central tendency will have a positive deviation; a score which is less will have a negative deviation. The sum of the deviations without regard to sign, divided by the total number of scores in the group, gives the *average deviation*, usually indicated by the abbreviation, A.D.

¹ For this method see King, W. I., *The Elements of Statistical Method*, pp. 134-37; or Thorndike, E. L., *Mental and Social Measurements*, pp. 44-46.

In calculating the average deviation of a group of scores, which are listed as being within an interval, it should be remembered that the scores are to be considered as being distributed evenly over the interval. Thus their average value is at the middle point of the interval. If the interval is from 12.00 to 12.99, the mid-point is 12.50, and in calculating the deviation each score in the interval from 12.00 to 12.99 should be treated as having a value of 12.50.

The method may be illustrated by calculating the average deviation of the distribution given in Table XXX. The

TABLE XXX. SHOWING THE CALCULATION OF THE AVERAGE DEVIATION

<i>Intervals</i>	<i>Frequency</i>	<i>Deviation</i>	<i>Deviation times Frequency</i>
30.0 to 30.9.....	1	12.3	12.3
28.0 to 29.9.....	1	10.3	10.3
26.0 to 27.9.....	3	8.3	24.9
24.0 to 25.9.....	4	6.3	25.2
22.0 to 23.9.....	2	4.3	8.6
20.0 to 21.9.....	4	2.3	9.2
18.0 to 19.9.....	3	.3	.9
16.0 to 17.9.....	5	-1.7	-8.5
14.0 to 15.9.....	6	-3.7	22.2
12.0 to 13.9.....	2	-5.7	11.4
10.0 to 11.9.....	2	-7.7	15.4
8.0 to 9.9.....			
6.0 to 7.9.....			
4.0 to 5.9.....	2	-18.7	37.4
2.0 to 3.9.....			
0.0 to 1.9.....			
Total.....	35		186.3

True median 18.7

$$\text{Average deviation } \frac{186.3}{35} = 5.3$$

median is 18.7. The average of each group of scores, subtracted from the median, gives the deviations. When there is more than one score in an interval the deviation must be multiplied by the frequency to obtain the total deviation. The calculation of the average deviation may be shortened by using the approximate median, and then correcting the result.¹

(2) Percentiles; quartiles; probable error. Another measure of the variability of scores may be found by determining the points on the scale between which the middle 50 per cent of the scores are included. The lower one of these two points is called the 25 percentile. It is the point below which there are 25 per cent of the scores. The upper one is called the 75 percentile. One half of the difference between these two points is called the quartile range, or simply Q. In case the median is used as the central tendency, the quartile range is the same as the median deviation or probable error (P.E.).

The calculation of the 25 percentile and the 75 percentile is similar to that of the median, which is sometimes called the 50 percentile. The only difference is that instead of the half sum, the one fourth sum and the three fourths sum are used.

For the distribution given in Table XXIX, the one fourth sum is 9. The approximate value of the 9th score is 14.0. The correction is 1.0, making the true value of the 25 percentile 15.0. The three fourths sum is 25. The approximate value of the 25th score is 22.0. The correction is 1.0, and the

¹ For this short method see Thorndike, E. L., *Mental and Social Measurements*, pp. 46-47.

true value of the 75 percentile is 23.0. The range of the middle 50 per cent of the scores is from 15.0 to 23.0, or 8 units. The value of Q is one half of 8 or 4 units. This means that 50 per cent of the scores contained in this distribution are contained within an interval of 4 units on either side of the median. The quartile range or Q is easy to calculate, and for many purposes is the most significant measure of variability.¹

Correlation. The Courtis Standard Research Tests, Series B, measure simultaneously two related factors of a pupil's ability to do the operations of arithmetic with integers, — speed or number of examples attempted, and the accuracy or per cent of examples done correctly. In the measurement of handwriting scores of speed and quality are obtained. In reading the pupil's rate of reading and his comprehension are measured.

When pairs of scores are to be arranged in distributions it is convenient to use the plan shown in Table XXXI. This table shows the tabulation of the handwriting scores for a fourth-grade class. This table is read as follows: four pupils wrote at a rate between 20 and 29 letters per minute. The handwriting of one of these pupils was judged to be quality 20, one 40, one 50, and one 60. The line at the bottom marked "total" gives the distribution of the scores for quality. The column at the right marked "total" gives the distribution of the scores for speed. The tabulation shows the relation between speed and quality. The relationship of these two quantities is not constant. Certain pupils

¹ For methods of calculating the mean square deviation the reader is referred to Thorndike, E. L., *Mental and Social Measurements*, pp. 47-58.

possess a high degree of both speed and accuracy. Others are low in one and high in another. This type of relationship is called correlation. The degree of correlation may be expressed by a coefficient of correlation.¹

TABLE XXXI. DISTRIBUTION OF SCORES IN HANDWRITING FOR A FOURTH GRADE SHOWING THE ARRANGEMENT OF TWO KINDS OF SCORES TO SHOW CORRELATION

<i>Speed or number of letters written in three minutes</i>	<i>Quality scores (Ayres scale)</i>								<i>Total for speed</i>
	20	30	40	50	60	70	80	90	
20 to 29.....	1	—	1	1	1	—	—	—	4
30 to 39.....	2	—	1	—	2	—	—	—	5
40 to 49.....	1	1	—	1	—	—	—	—	3
50 to 59.....	3	1	—	—	1	—	—	—	5
60 to 69.....	—	1	2	—	—	—	—	—	3
70 to 79.....	—	2	—	—	1	—	—	—	3
80 to 89.....	2	—	2	1	1	—	—	—	6
90 to 99.....	—	2	—	1	1	—	—	—	4
100 to 109.....	—	1	—	1	—	—	—	—	2
110 to 119.....	—	—	—	—	1	—	—	—	1
Total for quality.....	9	8	6	5	8	—	—	—	36

Median: — Quality, 42; Speed, 63.

Coefficient of correlation. If two sets of quantities are related in pairs so that when one is large the other is large, and when one is small the other is small, there is said to be a positive correlation. If there are no exceptions to this relation, the correlation is called perfect, and the coefficient is +1.00. If the two sets of quantities are related so that when one is large the other is small, the correlation is

¹ For the method of calculating the coefficient of correlation the reader is referred to E. L. Thorndike, *Mental and Social Measurements*, pp. 156-85.

negative, and a coefficient of -1.00 indicates perfect negative correlation. A coefficient of zero or near zero indicates no correlation, that is, no constant relationship exists between the two sets of quantities. In general, coefficients between +.30 and -.30 should be interpreted to mean that no significant correlation exists. However, coefficients of correlation must be interpreted with care.

For some purposes a distribution, such as is shown in Table XXXI, is more significant than the coefficient of correlation. The study of such an array will reveal the general relation between the two abilities which prevails. If the scores are grouped near the diagonal from one corner of the table to the opposite one, a significant correlation exists. Sometimes this is brought out better by using a graphic distribution on a chart, rather than a statistical table form. This is well shown in Fig. 15. The college marks were determined by giving 6 for A, 4 for B, 3 for C, 1 for D, and 0 for F, and then adding the marks for the courses. Thus, five A's would give a total of 30, five B's a total of 20, etc. Some correlation is shown, but it is not marked.

In the chapter on arithmetic it was stated that each type of example requires a different ability. This is shown by giving the same pupils two tests, each test being confined to a single type of example. When the two sets of scores are tabulated so as to show the degree of correlation existing between them, it has been found that little correlation exists. This means that in general the possession of the ability to do examples of a given type is no assurance that a pupil possesses an equal degree of ability to do examples of a different type.

Graphical representation. Graphical representation frequently aids one in interpreting a group of facts. The general principle involved is that numerical quantities are

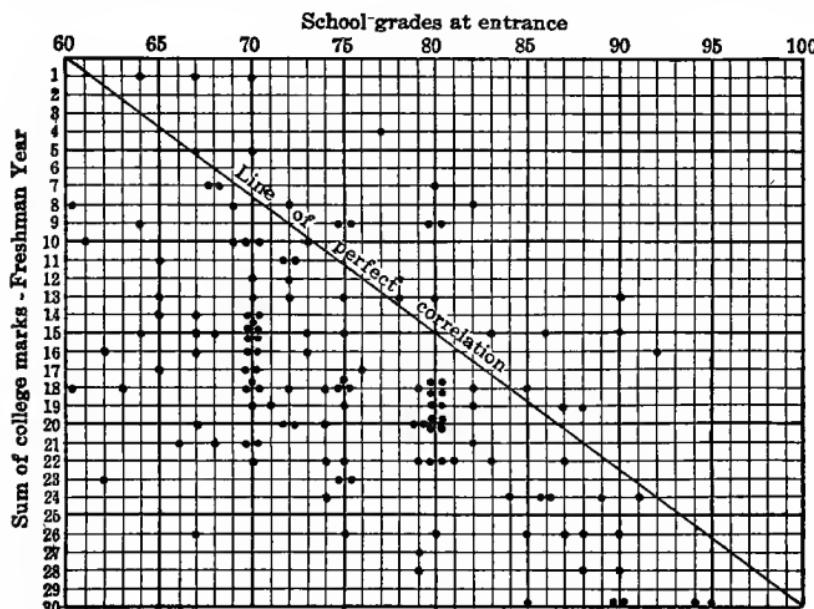


FIG. 15. SHOWING THE RELATION OF STANDING OF 138 COLLEGE STUDENTS, IN THEIR ENTRANCE EXAMINATIONS, TO MARKS MADE IN THEIR FRESHMAN YEAR IN COLLEGE.

(After Thorndike.) Each dot represents a student's record. The college marks were determined by giving 6 for A, 4 for B, 3 for C, 1 for D, and 0 for F, and then adding the marks for the courses. Thus, five A's would give a total of 30, five B's a total of 20, etc. Some correlation is shown, but this is not marked.

represented by the length of lines, distances from lines, or points, or areas, a certain length, distance or area having been chosen as a unit.

Certain types of quantities may be represented by a line. Take for example, the standards for Starch's Spelling Tests. (See page 130.) They are represented graphically in Fig. 16 by lines of appropriate lengths. This is often easier to grasp than a statistical form of statement. The meaning of the

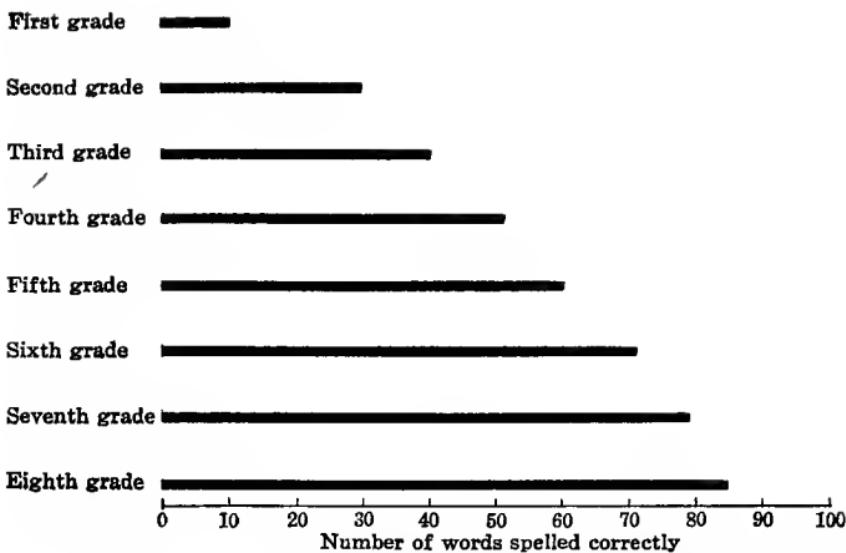


FIG. 16. A GRAPHICAL REPRESENTATION OF THE STANDARD SCORES FOR STARCH'S SPELLING TESTS. (See p. 130.)

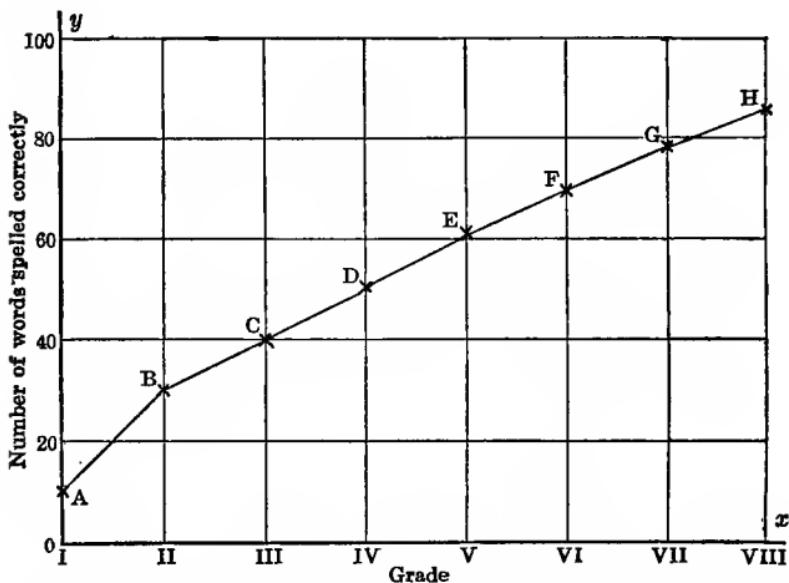


FIG. 17. ANOTHER FORM OF A GRAPHICAL REPRESENTATION OF THE SAME STANDARD SCORES AS IN FIG. 16

On this form of chart a number of pupil records may be dotted in. Those above the line are ahead of the standard for their grade; those below are behind.

length of the lines is ascertained by comparison with the scale at the bottom of the figure.

Another plan for representing graphically the same facts is given in Fig. 17. The base line $0x$ and the vertical line $0y$ are lines of reference. In drawing this type of graph, points are located whose perpendicular distances from the lines of reference represent the pairs of values. The distance

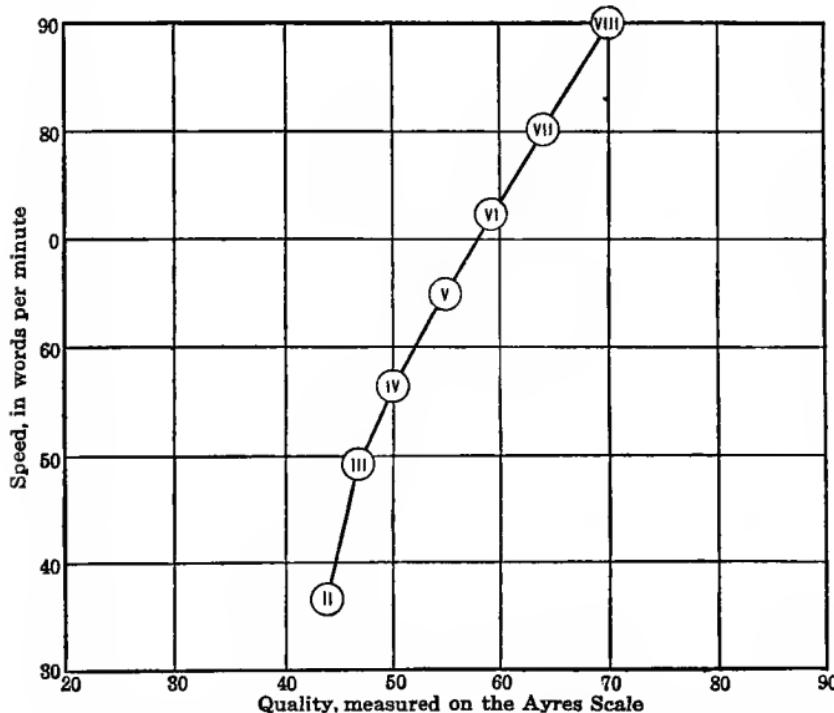


FIG. 18. REPRESENTING GRAPHICALLY THE STANDARDS FOR HANDWRITING

Three qualities are here represented — speed, quality, and school grade.

of the point C from the line $0x$ represents the number of words spelled correctly. The meaning of this distance is read from the scale on $0y$. In the same way the distance

of C from $0y$ represents the school grade. Thus, the position of the point C represents graphically the fact that the standard for the third grade is 40 words. The representation

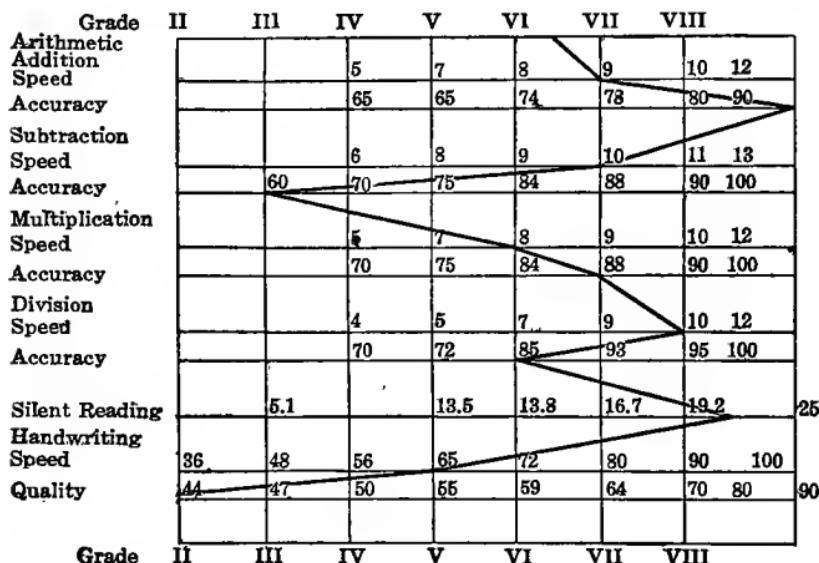


FIG. 19. REPRESENTING A PUPIL'S SCORES IN SEVERAL SUBJECTS

The figures on the chart represent scores or rates, and the position of the heavy irregular line the score x made by the pupil whose record is here charted.

stands out more clearly when the points are joined, as has been done in Fig. 17.

Representing three quantities. In the case of handwriting there are three quantities to represent — speed, quality, and school grade. The standards for the Ayres Scale (see page 168) may be represented as in Fig. 18. The speed and quality are both represented as well as the school grade. Roman numerals inside the small circle give the school grade to which the standards represented by the position of the circle belong.

Representing many quantities. Fig. 19 illustrates a plan

of graphical representation which is useful in representing the standing of a pupil or a class in several tests. The scales on the several lines have been so chosen that the standards for any grade lie on a vertical line. All of the standards for these tests for the fifth grade lie on the line marked V, those for the sixth grade on the line marked VI, and so on. The scores of a pupil or of a class are marked on the appropriate horizontal lines. When these points are connected, we have a graphic diagnosis of the ability of the pupil or of the general ability of the class.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is a distribution of scores?
2. What is a central tendency?
3. Calculate both the median and the average for a typical distribution. Compare the two central tendencies.
4. Do extreme scores affect the average? Do they affect the median?
5. Which of the three central tendencies is the most satisfactory for general use? Why?
6. Tabulate the handwriting scores of a class to show the correlation between speed and quality.
7. What are the advantages of graphical representation?
8. Why is a measure of the variability of a distribution needed in addition to the central tendency?

CHAPTER IX

THE MEANING OF SCORES

Indefiniteness of school marks. The result of measuring a physical object is expressed in terms of certain units, such as the foot, pound, square yard, or bushel. Likewise the result of measuring the abilities of pupils is expressed in terms of units. The speed at which a pupil is able to write is expressed in terms of letters per minute. The rate at which a pupil can add is expressed in terms of the ability to do a unit example. In the case of the Courtis Standard Research Tests, Series B, the unit addition example consists of three columns of nine figures each. The measure of an ability expressed in terms of a unit is usually called a score.

We apply certain descriptive terms to physical objects. By saying that a man is "tall" we express the fact that his height is greater than that of the average man. Since we know the height of the average man, a "tall" man is interpreted as one whose height is near six feet. However, a "tall" tree is not one which is six feet in height. A room eighteen by twenty feet would be a "large" kitchen, but a "small" classroom. What descriptive term or meaning is applied to an object depends upon a standard of size as well as upon the magnitude of the particular object.

School marks or "grades" are descriptive terms, similar to "tall," "large," "short," "heavy," and the like. Words such as "fair," "good," "excellent," "poor," "superior,"

and "medium" describe the attainments of pupils in comparison with a standard.

They do not express measures of pupils' abilities, but instead they are the meanings which teachers assign to measures of the abilities of pupils. To say that a fourth-grade pupil is a "superior" reader indicates that he possesses a higher degree of ability to read than the average fourth-grade pupil, but it does not tell how rapidly he reads, nor how well he comprehends. On the other hand, to say that this pupil reads a certain printed text at the rate of one hundred and sixty words per minute tells his rate of reading, but does not tell his standing as a fourth-grade pupil. His standing as a fourth-grade pupil can only be learned by comparing his reading rate with the standard rate for fourth-grade pupils.

School marks vs. scores. Usually a pupil's "grade" on an examination is the per cent of questions he answers correctly, credit being given for answers partly right. This practice means that the number and difficulty of the questions represent the standard of attainment for pupils of the grade to which they are given. A "grade" of 87 per cent placed upon an examination paper represents a comparison of the pupil's ability with the standard which the teacher established in the preparation of the examination. A low "grade" may be due either to the pupil's lack of ability, or to the high standard which the teacher set. Likewise a high "grade" may be due either to the pupil's exceptional ability, or to an easy examination. Thus "grades" expressed as per cents are not measures of the attainments of pupils, but only terms which describe pupils' attainments in com-

parison with standards established by the teacher. To define a "grade" of "excellent" as meaning from "95 to 100 per cent" is simply to exchange one descriptive term for another. It does not define "excellent."

Teachers frequently fail to distinguish between scores, and school marks or "grades." This is especially true when the examination is considered to consist of 100 points, and the "grades" are expressed in terms of per cents, but the two are not the same and should not be confused. The relation which exists between scores and "grades" may be shown by the following illustration. Suppose that a teacher of a seventh-grade class constructs an examination which she considers includes 100 units. In terms of the units of this examination, the five highest scores out of a class of 30 pupils are as follows: 80, 77, 75, 74, 72, and the five lowest are 49, 47, 46, 44, 41. Suppose, also, that in this school the school "grades" are to be reported in per cents, with 70 as the passing mark. Do these scores mean that only those pupils who have scores of 70 units or above are to be given passing "grades"? Certainly not. The scores in terms of units must be translated into school marks, but before this can be done, the basis of translation must be determined, and this basis involves knowing what scores seventh-grade pupils should make on this particular examination.

Translating scores into school marks. The conditions described above are represented graphically in Fig. 20. The base line marked from 24 to 100 represents the scale of the test, the portion from 24 to 0 being omitted because it is not used in the illustration. The five highest scores and the five lowest scores are marked on this scale. We know

that when a sufficiently large number of pupils are measured, with respect to any ability, they will be grouped in the manner indicated by either of the curves in the figure. There will be a few with very high scores, a few with very low scores, and a large number very near the average of the group. Knowing this fact, it is simply a question of where this distribution will be located on the scale line in the case of seventh-grade pupils.

The solid line curve indicates one possible location of the distribution. After the location of the distribution has been determined, there remains the placing of the passing mark in this location. This is an arbitrary matter, and any school may place it where it

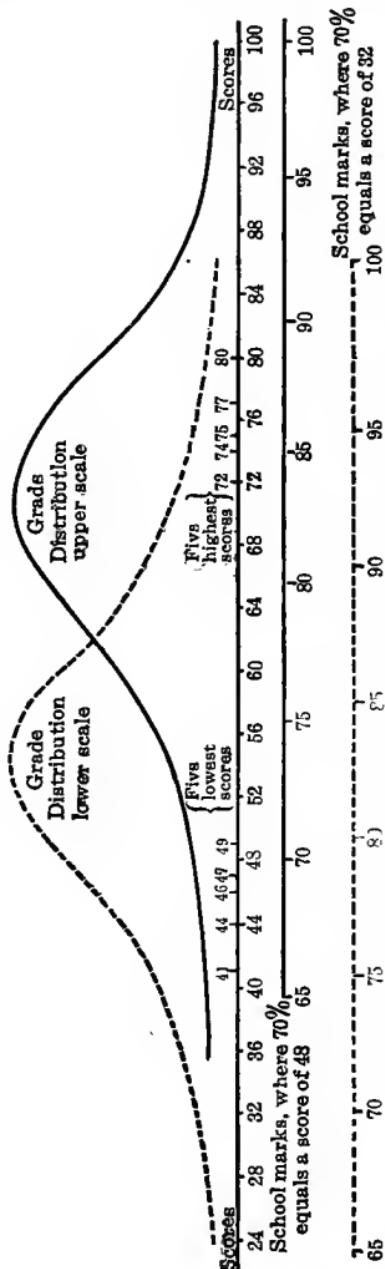


FIG. 20. SHOWING THE MEANING OF SCHOOL GRADES IN TERMS OF SCORES

Scores made by using standard tests are fixed and are equivalent anywhere and at any time; school grades are variable, and vary not only from school to school and teacher to teacher, but with the same teacher from month to month and day to day.

desires. Assume that in this particular case it is located at 48 on the scale of the test. Again this passing mark may be called anything as a school mark. Assume that in this particular case it has been called 70 in this school. Now, we have a satisfactory basis for translating the scores into school marks. The scores marked on the scale of the test are to be read from the solid line scale below. The pupil whose score was 49 would receive a school mark of 71. For only the four lowest scores a school mark of failure would be given. For the score of 80 a school mark of 88 would be given.

If, however, the location of the distribution of the scores of seventh-grade pupils should be that indicated by the broken line curve we would have a different basis of translation. The school marks would be read from the broken line scale. For the lowest score of 41 a school mark of 75 would be given.

This illustration should make clear that scores are not the same as school marks, and should also indicate the information which is necessary for a teacher to have before she can make an accurate translation of scores into school marks or "grades." Only in the case of tests which have been standardized will it be possible for the teacher to use such an elaborate plan of translation. Tests constructed by the teacher are not standardized, but a crude standard distribution may be obtained, as follows: Divide the test into convenient units or points. It is best to choose the unit, so that the total is not 100, but some other number as 28, 39, or 163. In reading the papers, assign scores in terms of this unit. If these scores are arranged in order

the teacher will have a crude standard distribution, and, by comparing individual scores with it, "grades" may be assigned more intelligently than by making no distinction between scores and "grades."

A satisfactory standard. If many of the critics of the public schools are only partly right in their contentions, present conditions are far from what they should be. Thus we are not warranted in assuming that standards which simply represent present conditions are satisfactory, and we may appropriately consider the basis of satisfactory standards.

A satisfactory standard must be reasonable and must be "efficient." To be reasonable a standard must be such that it can be attained by pupils, under school conditions, and with an appropriate time expenditure. Pupils are limited in their learning by inherited characteristics, and all cannot attain to the same levels of skill. However, it must not be forgotten that the medians, or averages, of present attainments of pupils are far below the levels of skill which have been attained by many pupils and adults. For example, the eighth-grade standard for the Courtis addition test is only eleven examples attempted. Some eighth-grade pupils are able to do twenty or more examples, and adults who have been specially trained have done fifty to sixty examples. In fact Courtis, who has studied arithmetical abilities for several years, states that it appears that there is no limit to the rate at which columns of figures may be added, provided the amount of time for practice is unlimited. In the school the time for practice is limited, but even then the standard of eleven examples is markedly below the attain-

ment of many pupils under present school conditions, and it still remains to be seen what degree of ability to perform the fundamental operations of arithmetic with integers might be engendered with the present time allotment, provided the methods and devices of instruction were properly suited to the pupils. It seems probable that the standards which we have now are below the level of the possible attainment of a large per cent of pupils under school conditions. Just how large this per cent will be when the methods and devices of instruction are appropriately adjusted to the pupils can be determined only by actual trial. It may be that, when we learn to adjust our methods and devices of instruction better to the abilities of pupils, a higher standard of ability may be attained with a less expenditure of time and teaching effort.

An efficient standard. The second qualification of a satisfactory standard is that it be "efficient." By this it is meant that the standard must represent a degree of ability which equips pupils for meeting present and future demands with a high degree of efficiency. The word efficiency has been borrowed or rather adopted from the field of engineering and mechanics. The efficiency of a machine such as a steam engine is the value of the fraction whose numerator is the amount of work which the engine does, or its accomplishment or output, and whose denominator is the amount of energy put into it in the form of fuel. The value of this fraction may be increased in two ways: first, if the numerator, that is, the amount of work done, is increased without increasing the amount of energy put into the engine, or at least without increasing it in the same proportion; second,

if the denominator is decreased without decreasing the numerator in the same proportion. The most efficient machine is the one for which the value of this fraction is the largest.

The word efficiency with essentially this meaning is now employed with reference to many forms of human endeavor. The numerator consists of the actual accomplishment. The denominator consists of energy and time which are put into the project, both in the form of preparation and in the actual doing at the time. For example, the contractor for a building erects a tower for hoisting and distributing the concrete used in the construction. He installs a rock crusher and other machines and appliances which will form no part of the completed building. All of these things are a part of the expense he puts into the building. In addition, he puts a large quantity of labor into it. These two items, together with the building material, are measurable in terms of dollars and cents and constitute the denominator of the fraction. The product of his endeavor, the completed building, forms the numerator of the fraction. His efficiency as a contractor is represented by the ratio of these two quantities.

He might have dispensed with the tower and have had the concrete distributed in wheelbarrows. He might even have dispensed with the rock crusher and mechanical mixer for the concrete. By so doing he would have eliminated these items of expense, but it is reasonably certain that if he had done so the total expense of constructing the building would have been considerably increased by the added labor, and this would have decreased the efficiency of the enterprise. The total expenditure of time and effort or money is the sum of the expenditures for preparatory and accessory

purposes, plus the expenditures of actual operation. In a large project, if the expenditure for preparatory and accessory purposes is too small, then the operation expenses are unduly large, making the total larger than necessary. A high degree of efficiency demands that there be such an adjustment between the two as to make the sum as small as possible.

Effort to be expended on the tool subjects. The school subjects are frequently classified under two heads:— tool subjects, and content subjects. The tool subjects of the elementary school are reading, handwriting, the operations of arithmetic, spelling, and language. In the study of the content subjects, such as the problems of arithmetic, literature, geography, history, science, etc., the tool subjects are used. The situation with respect to school subjects is quite analogous to that of the illustration just cited. The tool subjects are used in further learning in school, and in practical activities outside of school. Time and effort are required for acquiring skill in using these tools. Time and effort are also required when these tools are used. If only a small degree of skill is acquired, the time and effort required for using the tools are greatly increased. For example, time and effort are required to learn to add. By increasing the amount of time for practice, the skill of the learner can be increased and the time required to add numbers in the solving of problems and in the practical activities will be decreased. If the learner has a large amount of adding to do, it will be economy for him to spend a relatively large amount of time in practice; that is, in preparation. If he is going to have only a few occasions to add numbers, it will not be economy for him to spend a large amount of time in practice.

The situation is precisely the same as that of the contractor who was mentioned. When he is constructing a \$500,000 building it is economy of time and effort (both of which have a money equivalent) to spend several hundred dollars and several weeks of time in preparation for the construction of the building. If, however, he were building a \$3000 residence it would be folly to spend a very large amount in preparation for the work. So if the pupils in our schools are going to have many occasions to add in their future school work and in their activities outside of school, it will be economy to spend enough time and effort to engender in them a relatively high degree of skill. If, on the other hand, these pupils are going to have only a few occasions to add, it is folly to expend the time and effort to engender in them a high degree of skill. What is true of addition is true of the other operations in arithmetic and of the skills involved in the other tool subjects.

School demands on the tool subjects. Some of the occasions for the use of these tools occur in the work of the school, and some occur outside of school in practical activities. It is generally conceded that these tools should be acquired in the first six grades. In the seventh and eighth grades and the high school pupils have many demands made upon them for reading, writing, spelling, the operations of arithmetic, and expression by means of language, both oral and written. Our manner of carrying on school work by the use of textbooks and reference libraries makes the demand for reading very heavy. It is also our custom to require much written work, prepared outside of the recitation period, and in some subjects much written work during the recitation period.

This custom makes heavy demands for writing, spelling, and written expression. In arithmetic we expect pupils to learn to solve problems (not examples) by solving problems. In fact we require them to solve many problems, and the solving of problems requires arithmetical operations to be performed. In view of the fact that the school itself makes enormous demands upon its pupils for the use of these tools, it is folly not to prepare them adequately for these demands. It would be just as sensible for a contractor of a \$500,000 building to fail to provide a mechanical mixer for concrete as for a school to fail to prepare its pupils to read with an appropriate speed and quality of comprehension. In the case of the contractor, failure to provide appropriate machinery means that the concrete must be mixed by means of back-breaking and time-consuming labor. In the case of the school, failure to equip the pupils properly to read means that the numerous assignments which they will be asked to read will not only consume an enormous amount of time, but will also destroy interest in the school work because for them reading is a slow and difficult and hence a disagreeable process.

Outside of school there are a number of demands for these tools which are common to all: — reading newspapers, magazines, and books; writing letters; expressing ideas; and solving simple arithmetical problems of everyday life. In addition to these there are a number of special demands which depend upon one's occupation. Educators differ concerning the extent to which public schools should prepare pupils for these special demands, but practically all agree that little differentiation should be made below the sev-

enth grade, and, therefore, the question of preparation for these special demands concerns us but little in the consideration of standards for the tool subjects.

Basis for standards of accomplishment. The demands of the school, and the common demands of life outside of school, are the requirements which are to be considered in the setting up of standards for the tool subjects in the elementary school. In general discussions concerning what the schools should accomplish, and in practically all of the discussions of particular standards, attention has been focused upon the demands of life outside of school, and the demands of the school have been overlooked. This is perhaps due to the recent emphasis upon the fact that the function of the school is to give children preparation for the activities of life outside of school. This is a most wholesome and commendable point of view, but its acceptance should not blind one to the fact that the demands which the activities of the school make for the use of these tool subjects exceed many of the demands which the common activities outside of school make. The average man or woman does not meet as pressing demands for reading as do the pupils in the high school. Likewise the demands for writing, and probably for the other tool subjects as well, which pupils meet in school are greater than they will meet outside of school.

By saying that they are greater it is meant, not only that the demands are more numerous but also that they involve a limited time for their satisfaction. For example, when a pupil is given a test in school it is not only necessary that he write legibly, but it is also very necessary that he write

reasonably rapidly and without focusing his attention upon the act of writing. If he does not he is seriously hindered in answering the questions.

Even if the tool subjects were not practical, it would be necessary for the school to teach them and teach them well in the first six grades, in order that the pupils might do the work of the following grades. There is no valid basis for the argument that, since few of the pupils will become bookkeepers or clerks, or enter other specialized occupations, emphasis upon definite standards of skill in performing the operations of arithmetic and in handwriting is unjustifiable.

It should, however, be recognized that in the case of the content subjects the source of the standards of attainment is the demands of life outside of school. For example, if we were considering standards for the solving of problems in arithmetic instead of the doing of examples, the source of standards would be the demands which exist in the practical activities of life.

It should be evident from the foregoing discussion that a standard may be set too high. On the other hand it may be too low. Either condition means a low degree of efficiency. A teacher should not take pride in the fact that she has brought her pupils up to a point well above the standard. This condition may mean that she is just as inefficient as the teachers whose pupils are below standard, her inefficiency being due to an unusual expenditure of time for the engendering of this particular outcome.

Types of standards. The measures of a class or other group of pupils taken as a unit are represented by the central tendency and variability of the distribution of the

individual scores. In some classes we find a very great range of ability; in others it is very much less. The range of the distribution of the scores or the variability should be considered, as well as the central tendency. For giving meaning to these measures corresponding standards are required. In the absence of scientifically-derived standards the corresponding central tendency and variability of a large group of pupils are used. Another basis of interpretation is to use the corresponding measures of a number of similar groups. The position which a particular class occupies among a number of classes gives a meaning to its scores. The comparison of the scores of a class with those for classes in grades above and below is often illuminating, because the instruction of a school should be organized so that the pupils' progress through the several grades will be systematic.

Class scores are only indices of the general standing of the class. Since the teacher is instructing individual pupils, rather than a group of pupils taken as a unit, she needs to interpret the score of each pupil. Averages or medians alone are insufficient to do this because pupils differ widely in ability. To interpret fully individual scores it is necessary to have a standard distribution of scores. The meaning of a score depends upon its position in this standard distribution. The method of doing this is shown in the illustration on page 260.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What is the distinction between "scores" and "school marks"?
2. What things must be known in order to translate scores into school marks?

3. What do you think of the plan of placing scores on pupils' report cards instead of school marks? (It is assumed that if scores are used the standards will be given also.)
4. Why are school marks, such as "92 per cent," "good," "excellent," and the like, indefinite?
5. Why are standards necessary?
6. Why are averages and medians not satisfactory standards?
7. How must satisfactory standards be determined?
8. What is the meaning of "efficiency" in education?
9. How is a pupil's score interpreted?
10. What different types of standards are available?
11. What is a "standardized" test?

CHAPTER X

THE DERIVATION OF TESTS, AND EXAMINATIONS

Analyzing pupils' class work. In constructing an instrument to measure the achievements of pupils, the first step is to secure the exercises or specimens of pupils' productions which shall constitute the test or scale. Different school subjects present different problems, but in all cases this first step involves the analysis of the subject-matter, to determine the achievements which are fundamental. A brief account of the necessary analysis for certain subjects has been given in the preceding chapters, under the head of the "problem of measurement" and in the description of certain tests and scales. The method of analysis for the operations of arithmetic is well illustrated in the derivation of the tests for addition of fractions. (See page 34.) Upon the hypothesis that each type of example requires a specific ability, the addition of two fractions was analyzed to determine the different types of examples which might arise. This analysis was verified by a study of the ability of pupils to do the several types of examples.

In the field of spelling a different type of analysis has been made. By examining the writings of children and adults, Ayres and others have determined the most frequently used words. Starch analyzed the non-technical words of the English language so as to secure random samples for his spelling tests. In addition it has been shown that the ability to spell

words in dictated lists is not the same as the ability to spell the same words when they are given in timed sentences.

Exercises or specimens of pupils' productions are only the crude materials out of which a measuring instrument is to be constructed. The measurement of quantity requires a unit. For measuring length we have the yard and meter; for measuring weight, the pound and gram; for measuring time, the second. For measuring the abilities of pupils to perform the operations of arithmetic, to write, to read, to translate Latin, and the like, units of elemental abilities are necessary. The abilities of pupils are measured by having them do certain exercises, or by comparing specimens of their work with specimens of known value. The exercises or specimens of pupils' work which constitute a test or scale must be evaluated in terms of a unit.

Bases for evaluating pupils' work. The value of an exercise or specimen may be considered from the standpoint of its importance or of its difficulty. The time required in doing an exercise is a third factor, if difficulty is defined in terms of the per cent of correct answers. It is not possible to measure objectively the importance of an exercise. It is an easy matter to secure a measure of the difficulty of an exercise by having it given to a large number of pupils. The per cent of pupils who solve it correctly is an index of its difficulty. For this reason exercises are generally evaluated on the basis of their difficulty. Specimens of pupils' productions, as in the case of handwriting or English composition, obviously cannot be evaluated on the basis of difficulty of production. In constructing scales for these subjects,

specimens have been evaluated on the basis of the consensus of opinion of competent judges.

The nature of the subject-matter of arithmetic is such that it is possible to construct examples which are approximately equal in difficulty and hence represent equal degrees of ability. A multiplication example consisting of a four figure multiplicand and a three figure multiplier is approximately equal in difficulty to any other multiplication example similarly constructed. Hence, in arithmetic, each of the exercises which compose the test may be constructed equal in difficulty, thus avoiding the necessity of evaluating them. However, with the partial exception of algebra, this is not true of the other school subjects.

The cycle principle. In constructing the Standardized Algebra Tests, Rugg has employed the "cycle principle" of rotation.¹ This principle is employed when it is desired to include two or more types of examples in a single test. According to this principle the examples are arranged so that each type of example recurs at a regular interval. The cycle, instead of a single example, is the unit. Thus the necessity of evaluating the exercises in terms of a common unit is avoided.

The per-cent-of-pupils-solving basis. The most frequently used method² of evaluating the exercises of a test

¹ Rugg, H. O., and Clark, J. R. "Standardized Tests and the Improvement of Teaching in First-Year Algebra"; in *School Review*, vol. 25, pp. 113-32.

² A detailed description of this method, applied to a particular subject, is given in each of the following monographs:

Buckingham, B. R. *Spelling Ability; Its Measurement and Distribution*.

Trabue, M. R. *Completion-Test Language Scales*.

Woody, Clifford. *Measurements of Some Achievements in Arithmetic*.

is to have them given to a large number of pupils. The relative difficulty of exercises depends upon the per cent of responses which are correct, but the degree of difficulty is not proportional to the per cent of correct responses. For example, suppose exercise A is answered correctly by 40 per cent of a group of pupils, and exercise B by 80 per cent of the same pupils. Then exercise A is more difficult than B, but not necessarily twice as difficult. This is due to the fact that children do not exhibit uniform differences in ability.

A normal distribution of ability. It is a well-known fact that when a group of pupils is measured with respect to a mental or physical characteristic they are found to be distributed as shown in Fig. 21. The pupils near the median ability-group differ only slightly in ability, while those at either extreme exhibit large differences in ability. The degree of difficulty of an exercise which is done correctly by 80 per cent of the pupils is represented by the position of A

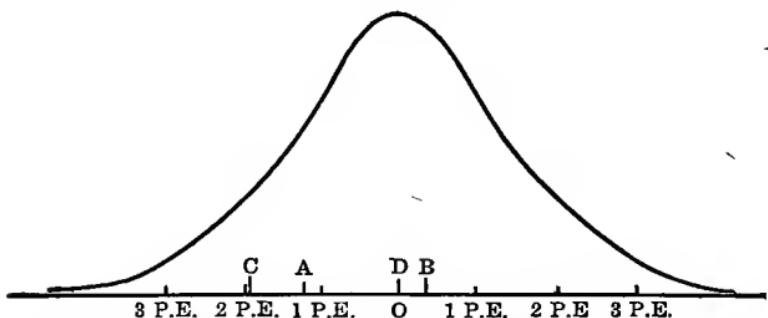


FIG. 21. SHOWING A NORMAL DISTRIBUTION OF PUPILS ACCORDING TO ABILITY

in Fig. 21. The degree of difficulty of an exercise, answered correctly by 40 per cent of the pupils is indicated by the position of B. The degree of difficulty of exercises which were

done correctly by 90 per cent and 50 per cent of the pupils is indicated by the position of C and D respectively. The difference in degree of difficulty which is represented by the difference between 90 and 80 per cent of correct responses is obviously not equal to the difference in degree of difficulty which corresponds to a difference between 50 and 40 per cent of correct responses. This illustrates the relation between the degree of difficulty of an exercise and the per cent of correct responses. Tables have been prepared to show the degrees of difficulty, corresponding to the various per cents of correct responses.¹ The degree of difficulty is expressed in terms of the probable error (P.E.). (See page 249 for definition.) The P.E. is used as the unit, since it is a constant function of all normal groups.

Points to be considered in evaluating exercises. In translating the per cent of correct responses into P.E. equivalents the values are counted from the median of the distribution. To express the absolute value of an exercise a zero point must be established. This is done by constructing an exercise which calls for zero ability. The other exercises are then compared with this one. For the method used see the work of Buckingham, Trabue, or Woody.

Exercises which are foreign to the experience of pupils do not furnish a satisfactory measure of their ability. Exercises which are suitable for eighth-grade pupils may not be appropriate for pupils in the fourth or fifth grade. In constructing a test which is to be used in several grades it is necessary to reject those exercises which are foreign to the experience of the pupils in any of those grades. In deriving

¹ See any one of the monographs referred to above.

his arithmetic scale Woody selected only those examples which were done correctly by a gradually increasing per cent of pupils in the grades from two to eight.

In evaluating the exercises of the Kansas Silent Reading Tests, Kelly considered the time taken by pupils in doing them, as well as the correctness of their answers. The average number of seconds required to produce a correct answer was obtained by dividing the total time used by the number of correct answers. The values of the exercises were made proportional to the average time required to produce a correct answer.

Opinion of competent judges. Hillegas used the consensus of opinion of "competent" judges in evaluating the merit of pupils' compositions, in constructing his composition scale. This method has been used also by Thorndike, in constructing scales for handwriting, drawing, and algebra. The method is based on the theory that differences of merit or quality which are noticed equally often are equal. Suppose a set of compositions, A, B, C, D, E, etc., are arranged in order by 100 competent judges. Suppose B is ranked better than A 75 times and poorer than A 25 times; suppose C is ranked better than B 75 times and poorer than B 25 times; suppose D is ranked better than C 75 times and poorer than C 25 times. Then the difference in merit between B and A is equal to the difference between C and B, and is equal to the difference between D and C. Or, in other words, the successive differences in merit between the four compositions A, B, C, and D are equal. By having a large number of specimens ranked in order by competent judges, it is possible to select a sufficient number of

specimens representing equal differences of merit to form a scale. Thorndike recommends that the difference which is noticed by 75 per cent of the judges be used as a unit.¹

The teacher-judgment basis. Ballou used the teacher-judgment basis in constructing the Harvard-Newton Composition Scale, but employed it in a different manner. Instead of having the compositions ranked singly in order of merit, he had numerical values assigned to each by each teacher, following certain directions. The average of these values was taken as the true value.

Reliability important. If a test or scale is to be effective as an instrument for educational diagnosis, it must yield reliable measures of definite abilities. The definiteness of the measures is determined by the selection of the exercises. If the exercises are restricted to a single type, as in the Courtis Standard Research Tests, Series B, the scores will have a definite meaning. The reliability of the scores depends upon the method of using the test or scale. Slight variations in the time allowed, or even in the manner the pupils are approached, affect the scores. Scores are also affected by the plan employed in marking test papers. If the exercises admit of only one correct answer, and if they are marked either right or wrong, no variation is possible. But if it is left to the teacher to judge the merit of the answer, as is the case in certain reading tests, there is certain to be a large degree of variation. Since scores are given meaning by comparing them with standards, it is necessary that the scores be ob-

¹ For details of the method see Thorndike, E. L. *Mental and Social Measurements*, pp. 122 and 124.

tained under the same conditions. Thus a vital feature of a test or scale is definite directions for using it.

The use of tests and scales must not be too complex or make unreasonable demands upon the teachers' time. Tests which can be administered to only one pupil at a time require too much time. The plan for scoring the test papers may be so complex that it requires an unreasonable amount of time. Individual scores must be tabulated to determine class scores. For this purpose tabulation sheets must be provided unless it is left to the teacher to construct her own.

Making examinations more effective. In Chapter I it was shown that the measures obtained by giving ordinary examinations were unreliable. It is possible for teachers to increase greatly the reliability of these measures. A systematic plan for marking examination papers will materially reduce this source of error. Kelly¹ describes the following experiment. Six fifth-grade teachers gave a uniform examination in arithmetic to their pupils. Each teacher marked the papers for her own pupils, but did not record the marks on the papers. The superintendent asked a teacher who was unusually systematic in marking examination papers, to prepare a definite plan for marking these papers. After she had done so, she marked all of the papers in accordance with this plan. Then the teachers who had first marked the papers marked them a second time following her plan. This provided two marks for each paper by the classroom teacher, the first without following a systematic plan, and the second using a definite plan. Each of these marks was compared with the mark of the teacher who

¹ Kelly, F. J. *Teachers' Marks*, p. 84.

marked all of the papers. In Table XXXII the six teachers are designated by the letters A, B, C, D, E, and F. The table is read as follows: when no systematic plan was followed, teacher A marked one paper 16 to 20 points lower than the "judge," one paper 7 points lower, two papers 4 points lower, two papers 2 points lower, agreed with the "judge" on one paper, etc. The differences between the marks given when the classroom teachers had no standard or systematic plan and when they followed a standard are very striking. In the first instance the marks assigned by the teachers agreed with those assigned by the "judge" in only 5.5 per cent of the cases, while in the second instance they agreed in 63.5 per cent of the cases.

Care in framing questions. Reliability in marking examination papers also can be increased by exercising care in constructing the questions. Frequently the same purpose can be realized by forming the question so that only one correct answer is possible.

The error due to the unequal value of the questions can be reduced. Obviously it is impossible for a teacher to employ the elaborate method described in this chapter for evaluating the questions, but approximate relative values can be determined from the ratio of the number of correct answers to the number of wrong answers. Even if this is done in only a very crude fashion, it is worth while.

The rate at which the pupil works should be recognized, particularly when the exercises call for automatic responses. When the pupil is required to reason out his answer, the time he takes is not so important, but probably should not be entirely neglected. It is very easy to measure the rate at

TABLE XXXII. DISTRIBUTIONS OF DIFFERENCES BETWEEN TWO TEACHERS' MARKS ON SETS OF FIFTH-GRADE ARITHMETIC PAPERS — FIRST, WITHOUT ANY EFFORT TO UNIFY THE METHODS USED, AND SECOND, BY A COMMON STANDARD (AFTER KELLY)

Range of Differences	Without standard						With standard						
	A	B	C	D	E	F	Total	A	B	C	D	E	F
21 or more.....	2	..	2
16 to 20.....	1	1	1	3
15.....	2	2
14.....	1	1
13.....	1	2	3
12.....	..	1	1	..	2
11.....	1	..	1	2	4	1	1
10.....	1	1	..	1	1
9.....	..	1	2	1	4
8.....	1	3	1	6
7.....	1	1	..	1	1	1	6	1	..	1
6.....	..	2	..	1	1	1	4
5.....	..	1	2	1	1	2	7
4.....	2	2	2	1	1	2	10	1	2
3.....	..	4	2	1	2	2	11	1	1	1	3
2.....	2	2	1	1	1	1	8	4	1	1	3	7	1
1.....	..	5	4	3	2	4	18	2	3	4	5	1	16
0.....	1	4	4	1	1	1	12	22	30	16	16	29	26
1.....	2	5	2	2	2	1	14	5	..	2	2	1	3
2.....	6	1	3	2	3	1	16	1	1	3	5
3.....	9	..	2	..	2	..	13	..	2	2	1	..	6
4.....	5	1	4	1	5	1	17	..	2	3	3	..	8
5.....	2	3	2	2	1	..	10	..	1	1	2	..	4
6.....	1	1	..	3	2	..	7	1	1	..	2
7.....	..	1	1	6	1	..	9
8.....	..	2	1	2	..	1	6
9.....	1	..	1	2	4
10.....	1	1	..	1	3
11.....	..	1	1	2
12.....	1	1	..	1	3
13.....	..	1	1	1	3
14.....	1	..	1	1
15.....	1	1	2	1	..	1
16 to 20.....	..	2	..	1	2
21 or more.....	1	3	1	..	5
Totals.....	35	41	35	36	39	33	219	35	41	35	36	39	219
Medians.....	+3	0	+1	+6	-1	-4	+1

which the pupil works as well as his accuracy. This may be done by timing each pupil while doing a uniform amount of work. It may also be done by having all pupils work at a test a given number of minutes. From the amount of work each pupil does his rate of working can be computed.

The indefiniteness of the meaning of examination marks can further be reduced by confining the questions to one or two topics, or to a small group of closely related topics. Sometimes, too, it will be advisable to give a series of short examinations, rather than a single long one.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What does the problem of measurement involve?
2. What problems are involved in the construction of a test?
3. What methods have been used in determining the value of exercises?
4. What are the objections to the method of "opinion of competent judges"?
5. How may examinations be made more accurate measuring instruments?
6. Do you think that examinations have functions other than that of measuring the abilities of pupils? If so, what are they?
7. Repeat the experiment described on page 280 and compare your results with those given in Table XXXII.

CHAPTER XI

USE OF STANDARDIZED TESTS IN THE SUPERVISION OF INSTRUCTION

IN the preceding chapters standardized tests and scales have been considered primarily from the point of view of their usefulness to the teacher. They are also valuable to the supervisor. It is the problem of this chapter to set forth the supervisor's relation to the use of standardized tests and scales, and how they may be used by him in the supervising of instruction.

Assisting teachers. The supervisor (superintendent, principal or special supervisor) must assume the responsibility of educating the teachers of a school system in the making and using of educational measurements. This phase of educational development is so new that few teachers have become acquainted with it in the course of their professional training. It is important that the teacher think of tests and scales as instruments which will enable her to make her instructional efforts more effective. There is considerable evidence to show that if a teacher misinterprets the function of tests or looks upon them with suspicion, her efficiency as an instructor will be lowered by their use.

Some supervisors claim that the scores obtained by the use of standardized tests are an index of the teacher's efficiency. If the scores of her pupils are below standard, the teacher's efficiency is low. If they are above standard, the teacher's efficiency is high. Under certain conditions

this is probably true, but under the conditions which frequently prevail this conclusion cannot be justified. The efficiency of a teacher must be judged upon the basis of the growth which pupils make under her instruction. The low scores of a class at a given time may be due to the inefficiency of their present teacher or of their former teacher. Also, due consideration must be given to the quality of supervision, the specifications under which the teacher is working, the equipment with which she is supplied, and the native ability of the pupils.

The supervisor should assume the initiative in selecting suitable tests and scales and in planning their use. The criteria which should guide the selection have been given on page 85. Measurement at the beginning of the school year furnishes an inventory of the situation which the teachers and the supervisor face. Measurement at the close of the school year will reveal the extent to which they have met the requirements of the situation. In addition, measurement in September provides a check upon the work of the previous year. If a teacher has secured high scores in May by unfair or unwise means, this fact will be revealed when the pupils are tested again in September.

Four steps in supervising instruction with tests. In the making and using of educational measurements four steps may be recognized. First, giving the tests; second, tabulating the scores and calculating the central tendencies, variabilities, etc.; third, interpreting the scores; fourth, modifying instruction to meet the needs revealed. The supervisor can render valuable service in each of these steps. In a number of cities this service has been considered

important enough to justify the creation of a department of "educational research and efficiency." This has not been confined merely to the larger cities, but has been done in a few cities of 20,000 to 25,000. In a few cases similar departments have been created by county superintendents.

Giving the tests. The manner in which the test is presented to the pupils affects the scores. The purpose of measurement is defeated if the test is presented to the pupils in such a way that their response is unnatural. For example, in handwriting, if the pupils write at an unnatural speed the quality of their handwriting will be affected. It is also necessary that there be uniformity in making the measurements if comparisons are to be made. The supervisor should train the teachers to give the tests properly, or arrange to have some trained person give them to all classes.

Tabulating the scores. There must be uniformity in the tabulation of the scores. If directions and blanks for tabulation are not furnished with the tests chosen, the supervisor should decide what directions are to be followed. If the tabulation requires much time it is probably unwise to require the teachers to do it. Substitute teachers, and in some cases normal-training students, can be utilized for this purpose if trained clerical help is not available. In addition to the tabulations which may be made by the teachers, the supervisor should make others to show the situation for the school as a whole. Such tabulations are helpful not only to the supervisor, but also to the teacher. She needs to see her work in relation to the work of the school as a whole.

Interpretation of the scores. The interpretation of scores has been treated in Chapter IX, and hence can be passed over

with a bare mention here. The supervisor should assume responsibility for making the teachers aware of the complete significance of the information which the tests have provided. It is his function to provide standards and similar scores from other cities, if they are not easily accessible to the teachers. The significance of a group of facts is more easily grasped when they are represented graphically. The supervisor can render valuable service to the teachers by preparing a chart or series of charts to show the standing of the several grades of the school in comparison with the standards. Several schemes of graphical representation have been illustrated in Chapter VIII.

Remedial treatment. The fourth step, modifying instruction to meet the needs revealed, is the culmination of the first three. Without this step standardized tests become mere "playthings" and their use cannot be justified. The omission of this step creates a situation similar to that which would exist if a physician examined a patient carefully and determined the nature of his ailment but did not prescribe any remedial treatment. In our zeal to convert teachers to the acceptance of the principle that the measurement of certain results of instruction is possible there has been a tendency to overlook this step. In fact some have even said that they were content to apply the tests and reveal to the teachers the shortcomings of their work. These persons would leave to the teachers the difficult problem of remedying the defects. As a result not a few teachers have failed to see in the tests anything more than a new "plaything," which they might use to secure material for a paper to read at a teachers association or to arouse the

interest of their pupils. Such teachers have expressed their approval of the tests when their pupils' scores were high, and have considered the tests unsatisfactory when the scores were low.

Standardized tests are not "playthings." Neither are they teaching devices. They are instruments whose function is to reveal the conditions which exist so that the teacher's efforts to instruct her pupils can be made more effective. This function is not fulfilled until this fourth step is taken. In securing this fulfillment the supervisor renders significant assistance to his teachers. He can suggest remedial methods and devices. Numerous remedial devices have been mentioned in Chapters II to VII. Another device which has been tried with at least some measure of success is to make a special classification of the pupils on the basis of their ability, as shown by a standardized test, for the purpose of instruction in subjects such as spelling or handwriting.¹

Teachers need detailed and definite specifications. Detailed and definite specifications of the product are necessary for efficient work in even simple undertakings. It is particularly important in the education of children. Providing these specifications is the duty of the supervisor. Education consists of producing changes in children. When they enter the school they represent "raw" material which is to be modified by a series of workers (teachers). By means of text-books, and the other material equipment, and by means of assignments, questions, and directions, teachers cause children to acquire habits, knowledge, ideals, and other less

¹ Haggerty, M. E. "Some Uses of Educational Measurements"; in *School and Society*, vol. 4, pp. 761-71.

tangible controls of conduct. In the production of material things it is a cardinal principle that, "When the material which is acted upon by the labor processes passes through a number of progressive stages on its way from the raw material to the ultimate product, definite qualitative and quantitative standards must be determined for the product at each of these stages."¹

This principle applies also to the school. If the school is conducted in an efficient manner it is necessary that each teacher know in detail: (1) the changes which have already been produced in the partly educated children that come to her, and (2) precisely what changes she is expected to contribute to their education. Otherwise no teacher can proceed with her work in an intelligent manner. For example, the teacher of reading needs to know what rate of reading and what degree of comprehension she can expect of her pupils in September, and exactly what the school expects her to contribute to the reading ability of these pupils by June.

Courses of study represent working specifications. Our present courses of study represent the efforts of those occupying supervisory positions in our school systems to provide teachers with specifications for their work. How well they have succeeded is illustrated by the following quotations from typical courses of study.

FOURTH GRADE

Reading and literature. Stories read and told to the class; Roman stories, American history stories relating to geography, selections from Greek and Teutonic mythology, and poems.

¹ Bobbitt, Franklin. "The Supervision of City Schools"; in *Twelfth Yearbook of the National Society for the Study of Education*, part I.

A few choice selections of appropriate prose and poetry are to be studied, committed to memory, and recited or dramatized. See that the pupil stands on both feet and reads smoothly and confidently. Watch the voice of pupils; use breathing exercises; and avoid harsh, strained reading. Have pupils read many selections silently, then reproduce the thought aloud in order to develop the power of gaining and expressing the thought of the text. Aim to enlarge the pupil's vocabulary, to help him master the thought content, and gain the power to read in a pleasing, well-modulated tone. Use much supplemental reading. Explain the purpose of the children's department in the Public Library, and encourage pupils to read books therefrom.

Following these general directions, the selections to be read are specified.

THIRD GRADE: B CLASS

Handwriting. Daily drill, lessons five, six, or seven.¹ During the entire year these drills should be used for a few moments at the beginning of each writing lesson.

Beginning with lesson five, take the lessons in consecutive order to lesson thirty-five.

After developing a letter with the class take, from the writing book, a word beginning with the same letter and use it for practice.

If the letter is a capital, follow the word practice by using a sentence beginning with this letter. All words and sentences should be taken from the writing book.

GRADE 4 B. ARITHMETIC

Leading topics. The four fundamental processes with emphasis upon multiplication.

Review. Regularly, constantly, and from the first. The addition combinations, subtraction, reading and writing numbers, simple fractions.

Multiplication. The tables completed and made automatic. Problems with two-place multipliers. Rapid oral practice.

Division. Short division with long division brace. Rapid oral practice.

¹ Reference to a *Manual for Teachers*, used in this school system.

Fractions. Simple fractions and mixed numbers as needed in actual practice on concrete form problems. Largely oral and objective.

Concrete problems. One-step problems.

Applied problems. Farm products, farm marketing, farm profits.

Measures. Quart, gallon, peck, bushel, pound, ton, cord, etc., as required by applied problems.

Such subject-matter directions not quantitative. These specifications are in terms of subject-matter rather than results. Subject-matter is a way of acting.¹ A selection to be read is a requirement for a certain activity on the part of the pupil. Examples and problems in arithmetic are challenges to action. Drill exercises in handwriting call for motor activity. This activity is not the end to be attained but the changes which are produced in the child by his activity. Questions, drills, assignments, and the like serve to spur the child to action. As a result of his activity, motor, intellectual, or emotional, the child acquires habits, knowledge, and ideals. These are the ends for which the school exists.

Children differ in many ways. They differ widely in the amount of action which is necessary in order that they may acquire a given change. This fact is evident in any classroom. Under our present methods of teaching the same work is required of all pupils, and as a result we find that the pupils of any grade differ widely in their ability to do. One pupil learns to add twelve examples in eight minutes with 90 per cent of accuracy. Another pupil who has been subjected to the same training adds only seven examples in

¹ See Charters, W. W. *Methods of Teaching*, chapter II, for an elaboration of this definition of subject-matter.

this time, and only four are correct. One pupil learns to write at the rate of eighty-five letters per minute, and with a quality of 80 on the Ayres Scale. With the same training another pupil is able to write only sixty letters per minute, and with a quality of 50. Efficient teaching requires differentiated instruction rather than the same instruction for all pupils. This has been emphasized in the preceding chapters under the head of "remedial instruction." By stating the directions to the teacher in terms of subject-matter it is suggested that all pupils be given the same training, regardless of whether they need it or not. Since pupils differ widely, their needs cannot be the same.

Such directions lead to formal and uniform instruction. This emphasis upon subject-matter leads to formalism. Because the teacher is told to use certain subject-matter rather than to accomplish specified results, the use of subject-matter becomes her purpose. In teaching reading her purpose is to have the pupils read the required selections and books. In teaching arithmetic it is to have the pupils do the examples and problems on certain pages. In teaching geography it is to cover the specified topics. In manual training it is to complete the designated projects. The doing of these things has no importance except as the pupils are educated by reason of the activity. By transferring the purpose of the teacher, and through her the purpose of the pupils, from the end to be attained by the use of subject-matter to the use of subject-matter itself, the work of the school becomes formal.

When the ends to be attained are mentioned in courses of study the terms used are generally indefinite. "The mul-

tiplication tables completed and made automatic" is indefinite because there are many degrees of automatization. The degree of ability that one teacher would call "automatic" would not be accepted by another. "Rapid oral practice" is likewise indefinite. A definite statement could be made easily by specifying the rate in terms of responses per minute.

The tests aim to introduce quantitative work. The standards for scientifically devised tests define the achievements of pupils which are to be attained by the use of subject-matter. They make possible the writing of a course of study in terms of the results to be attained at each stage of the pupil's progress. For example, in handwriting the rate at which the pupil is expected to write and the quality of his writing are specified for each grade by the standards given in Chapter V. The abilities which the pupil is expected to exhibit in performing the operations of arithmetic are defined by the standards given in Chapter II. When the specifications for each stage of the work are expressed in terms of established standards the teacher knows what to expect of the children who come to her at the beginning of the year, and also what she is to contribute to their education. With her attention directed to the results to be secured rather than to the subject-matter to be used, the teacher has an opportunity to exercise her resourcefulness in using subject-matter as a means to that end.

Ability to do automatically is specific. The ability to spell "mountain" is distinct from the ability to spell "success." The ability to add a column of four figures is not the same as the ability to add a column of fifteen figures. The ability to add two fractions with unlike denominators is

not the same as the ability to add fractions with common denominators. A multiplicity of abilities must be engendered. The teacher must be made conscious of each ability as an end to be attained by each pupil under her tuition. Since the teacher is at all times concerned with the details of teaching, the general aims of education are not sufficient.

The inadequacy of general aims of education is well illustrated by the fact that frequently a teacher fails to recognize the existence of certain important details. Recently the writer found the handwriting of a certain supervisor of handwriting very difficult to read. An analysis showed that this was due to the lack of sufficient spacing between words. Another supervisor admitted that she had never thought of speed of handwriting as a factor in the aim of the teacher. Many teachers give evidence that their aim of teaching reading includes only that of oral reading.

Tests introduce scientific management. The principle of scientific management stated at the beginning of this chapter implies the material being acted upon must be tested at regular intervals in order to ascertain if the specifications are being met. By using standardized tests at regular intervals it is possible for a supervisor (superintendent, principal, or special supervisor) to know how teachers are meeting the specifications. This information is also valuable to the teacher. The use of standardized tests is most effective when teachers sympathetically coöperate with the supervisor in using them. The first effort of the supervisor should be to secure this coöperation. Independent use of tests by the teachers does not yield complete returns in increasing the efficiency of the school.

The time given to a school subject has a value in dollars and cents. Economy demands that no more time be given to a subject than is necessary for the pupils to attain a satisfactory standard of achievement. In the absence of definite standards school time has been allotted to the several subjects according to the opinion and interests of the supervisor. As a result the amount of time given to the several subjects varies widely.¹

Handwriting an example of wasting time. Investigation has shown that no relation exists between the time expended and the results obtained. The condition for handwriting in forty-seven cities is shown in Fig. 22. Similar conditions have been shown to exist in other subjects.

The comparison in the rank of schools which spend different amounts of time upon writing is shown in Fig. 22. Each vertical line in this figure represents one city. The lines upon the same horizontal line represent the cities which spend the same amount of time in writing. Those on the upper line spend the least amount of time, and those upon the lowest horizontal line, the largest. The position of the lines in the right or left direction represents the rank which was obtained by the schools as a result of the test. Those which are at the left side of the figure are higher in rank, and those which are toward the right are lower.

If the spending of a large amount of time in writing produces a corresponding gain in efficiency, the vertical lines should be grouped along a diagonal line running from the lower left-hand corner to the upper right-hand corner. That is, those which spend the less amount of time should be toward the right, and vice versa. It is evident that this situation is not represented by the facts. The cities which spend the various amounts of time are scattered throughout the range. For example, of the two cities which spend on the average only 45 minutes per week, one has the eleventh rank and the other the twenty-sixth; while two of the cities which

¹ See *Fourteenth Yearbook of the National Society for the Study of Education*, part 1, p. 26.

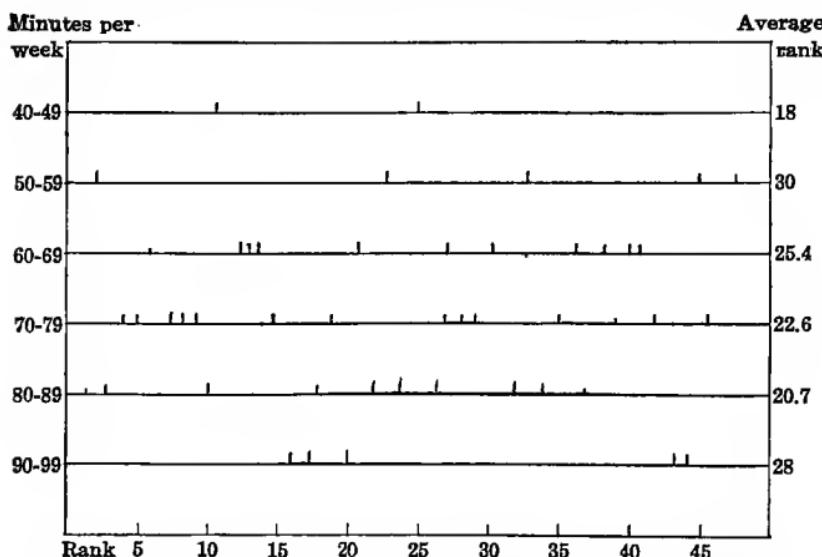


FIG. 22. DISTRIBUTION IN RANK OF 47 CITIES, ARRANGED IN CLASSES ACCORDING TO THE TIME SPENT ON HANDWRITING

spend an average of 95 minutes have the rank of forty-three and forty-four, very nearly at the bottom of the list. The average rank attained by the cities of each time-group are represented in the column to the right. It will be seen that, with the exception of the shortest-time and the longest-time groups, there is some increase in efficiency with an increase in time, but this increase, which holds on the average, is slight, and the exceptions are so great that the amount of time spent appears to have little influence upon the results.¹

Standardized tests furnish a means for the supervisor to determine in a scientific manner the optimum time to be given to each of the subjects.

The Cleveland reading results a study in efficiency. If the school is efficient a pupil's progress through the several grades must be in accord with a general plan. When the

¹ Freeman, F. N. *Fourteenth Yearbook of the National Society for the Study of Education*, part I, pp. 67-68.

teachers are working independently the pupil's progress from grade to grade may be very erratic. Figs. 23 and 24 show the conditions which were found to exist in silent reading in Cleveland, Ohio. These conditions are merely typical of many which measurement has revealed. The average scores

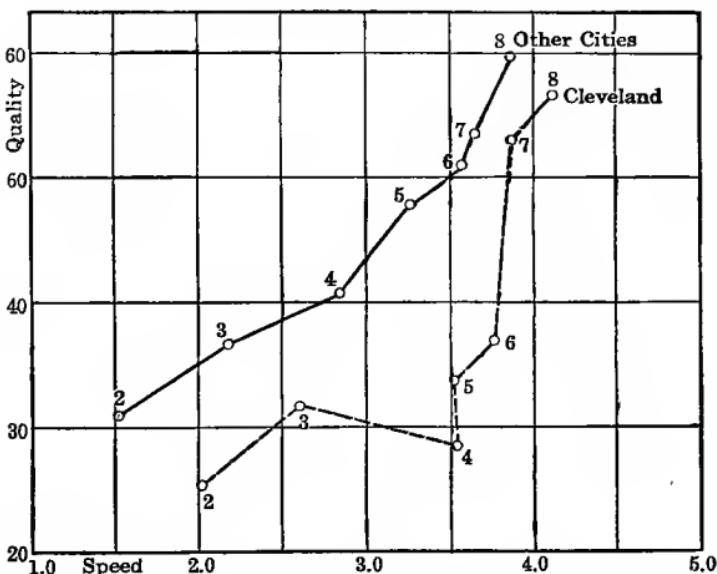


FIG. 23. AVERAGE SCORES IN SPEED AND QUALITY OF SILENT READING IN EACH GRADE IN CLEVELAND AND IN 13 OTHER CITIES. GRAY'S SILENT READING TESTS USED.

(From *Measuring the Work of the Public Schools*, by C. H. Judd.)

are indicated by the positions of the small circles. The numerals near the circles indicate the grades to which the scores belong. If the average scores of thirteen other cities are taken as a standard, the lack of satisfactory standards in certain grades is indicated for Cleveland. The progress from the third grade to the fourth is particularly unsatisfactory. However, Fig. 24 gives the most striking evidence of the lack of adequate supervision of the instruction in reading.

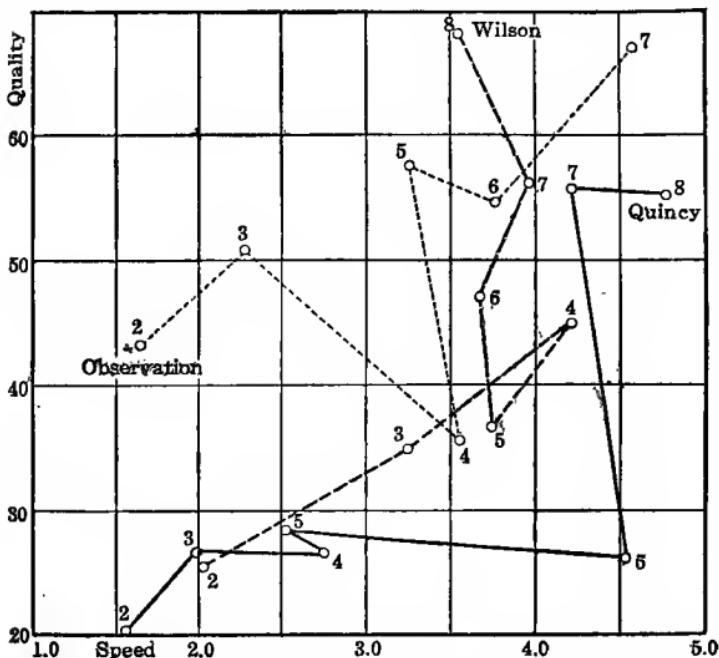


FIG. 24. SILENT READING SCORES IN THREE SELECTED CLEVELAND SCHOOLS

(From *Measuring the Work of the Public Schools*, by C. H. Judd.)

In any one of the three schools represented the progress of the pupils is so erratic that it is obvious that no definite plan existed.

Standards for instruction illustrated from arithmetic. What can be accomplished by the setting of definite standards for each stage of the pupil's progress and by systematic testing is illustrated by the following report:¹

The Courtis Standard Research Tests, Series B, were given in October, 1913, to the pupils in the four schools comprising a school system. The condition revealed was very unsatisfactory. The median scores were conspicuously below the standards for these

¹ Lane, Henry A. "Standard Tests as an Aid to Supervision"; in *Elementary School Journal*, vol. 15, pp. 378-86.

tests in practically all instances. The progress from grade to grade was irregular. The superintendent reached the conclusion that the inefficiency of arithmetic instruction was not due to faulty method so much as to the fact that teachers and pupils did not know definitely what was expected of them. He therefore placed in the hands of each teacher a copy of the following announcement:—

In June there will be another Courtis test of the same type as the one given in October. At that time, the grades are expected to attain the following standards:

	<i>Addition and Subtraction</i>		<i>Multiplication and Division</i>	
	<i>Attempts</i>	<i>Rights</i>	<i>Attempts</i>	<i>Rights</i>
4B.....	5.0	2.0	4.0	1.6
4A.....	5.7	2.9	4.9	2.4
5B.....	6.5	3.8	5.8	3.4
5A.....	7.3	4.7	6.6	4.4
6B.....	8.0	5.6	7.5	5.3
6A.....	8.8	6.5	8.4	6.2
7B.....	9.6	7.4	9.3	7.2
7A.....	10.4	8.2	10.1	8.1
8B.....	11.2	9.1	11.0	9.0
8A.....	12.0	10.0	12.0	10.0

Work in the four operations must be stressed. These are tentative and minimum standards. They will very likely be raised next fall. The success with which classes achieve these standards will, in a sense, be a measure of teaching ability.

When the tests were repeated at the close of the school it was found that "with very few exceptions the standards set were attained, and in some cases they were exceeded."¹

Results of using the Courtis tests in Boston. In certain cities tests have been used long enough to show the effect

¹ In using this illustration the writer assumes no responsibility for the standards used, or for the superintendent failing to realize the possibility that these standards might be lowered rather than raised.

of systematic measurement. In Boston the Courtis Standard Research Tests have been used in 29 schools since 1912. In the following statement these are called group A schools. Seventeen schools, Group B, have used the tests for one to two years. In seventeen other schools, Group C, the tests were given for the first time in May, 1915. In comparing the achievements of the pupils in these three groups of schools on the basis of the scores made in May, 1915, Ballou says:—

1. In the amount of work done by pupils in the four fundamental operations, Group A schools show superiority over Group B schools in sixteen out of twenty comparisons, and over Group C schools in eighteen out of twenty comparisons.
2. In the accuracy with which work was done, Group A schools show superiority over both Group B and Group C schools in seventeen out of twenty comparisons.¹

The supervisor and the standardized tests. Methods and devices of instruction must be judged by results. The supervisor can also use standardized tests in guiding teachers in the determination of the best methods and devices of instruction. When a test is given at the middle or close of the year the effectiveness of the instruction is indicated by the results. It will frequently happen that certain teachers are securing superior results. This is an indication that these teachers are using methods or devices of instruction which are superior to those employed by the other teachers. It is also possible to evaluate scientifically proposed methods and devices of instruction.

Formerly the duties and responsibilities of the supervisory officials were limited, for the most part, to enforcing disci-

¹ Ballou, Frank W. "Improving Instruction through Educational Measurement"; in *Educational Administration and Supervision*, vol. 2, pp. 354-67.

pline and to performing the clerical duties of their offices. A somewhat incidental duty was the supervision of instruction. Usually this meant assisting the inexperienced and less capable teachers. This was accomplished in two ways. The principal or superintendent visited the classroom and made note of such deficiencies as he thought important. These items were discussed with the teacher, remedies for the shortcomings being suggested. Another method was to take charge of the class and show the teacher how the work was to be done. This supervision was necessarily personal. It was the opinion of the supervisor against the opinion of the teacher. Standardized tests make possible a different type of supervision of instruction. By their use supervision becomes impersonal, it is no longer the opinion of the supervisor against the teacher. Both must submit to facts. By the use of standardized tests the supervision may be made scientific and one of the most significant results will be the development of a scientific attitude on the part of the teacher towards her work.¹

Standardized tests provide a means whereby a supervisor may render an accounting to the citizens of the community. If the business men question the quality of the product of the schools, he can present facts to show the quality as compared with that of other cities. Superintendents have testified that standardized tests would be worth while if they did nothing more than furnish an effective reply to the chronic faultfinders.

¹ Read in this connection Morrison, J. Cayce. "The Supervisor's Use of Standard Tests of Efficiency"; in *Elementary School Journal*, vol. 17, pp. 335-54.

But standardized tests do more. Their use eliminates "guesswork." The supervisor may now think about his work in the same type of terms as the manufacturer uses. He now has objectively defined units of achievement which he may use. The interpretation to be placed upon a given degree of achievement is not a matter of personal opinion. Standards furnish an impersonal basis which must be recognized by all. The value of being able to think about one's work in terms of facts and objective units, instead of in terms of opinions and vague terms, may not be immediately apparent. In the business world it is represented by the difference between success and failure. It is difficult to conceive any reason why the same significance should not exist in the field of education.

QUESTIONS AND TOPICS FOR INVESTIGATION

1. What are the steps in supervising instruction with standardized tests?
2. Should teachers use standardized tests if the fourth step is not taken? Why?
3. Why must the supervisor assist the teacher in this work?
4. Why is a general aim not sufficient?
5. What are the objections to our present courses of study with respect to the statement of aim?
6. How can standardized tests be used in setting the aim for the teacher?
7. Summarize the uses which a supervisor can make of standardized tests.

INDEX

INDEX

- Ability**, relation to performance, 21.
- Algebra**, problem of measurement in, 224; fundamental operations of, 225; remedial instruction in, 231 ff.
- Algebra Tests**: Standard Research Tests in Algebra, 228, 231; Thorndike's Algebra Tests, 228; Indiana Algebra Tests, 229, 239; Standardized Tests in First-year Algebra, 229, 239.
- Analysis** of ability in arithmetic, 19 ff., 37.
- Anderson**, H. W., 111.
- Arithmetic**: problem of measurement, 17 ff.; types of examples, 19; laws of habit formation applied, 53 ff.; remedial instruction, 53 ff.; individual differences, 55 ff.; devices for remedial instruction, 59 ff.
- Arithmetical Tests**: Courtis Standard Research Tests, Series B, 23, 38; Cleveland Survey Tests, 25, 40; Woody Arithmetic Scales, 29, 41; addition of fractions, 34, 42; Stone's Reasoning Test, 35, 42; Starch's Reasoning Test, 37; Courtis's Reasoning Tests, 37.
- Arithmetical abilities, nature of, 18 ff.
- Ashbaugh**, E. J., 171, 172, 173.
- Average**, 247.
- Average deviation**, 247.
- Ayres**, L. P., 15, 113, 116, 126, 148, 152, 156, 173.
- Ayres's Handwriting Scale, 152; Adult Scale, 152; "Gettysburg Edition," 152.
- Ayres's Spelling Scale, 113-16; 121.
- Ballou**, F. W., 18, 34, 40, 194, 279, 300.
- Bell, J. C., 233.
- Bobbitt**, Franklin, 289.
- Boston, standard scores for Courtis Standard Research Tests, Series B, 40; standard scores for addition of fractions, 42; results of using Courtis Tests, 299; copying test, 215.
- Breed, F. S., 164, 194.
- Breed and Down's Handwriting Scale, 153.
- Breed and Frostic Composition Scale, 195.
- Brown**, H. A., 74.
- Brown's Silent Reading Test**, 74-76; 87-89.
- Brownell, Baker**, 194, 200.
- Buckingham**, B. R., 115, 275.
- Buckingham's Spelling Scale**, 124.
- Buckingham's Grammar Test**, 215.
- Butte, Montana, scores for Stone's Reasoning Test, 43-44.
- Carter**, R. E., 3.
- Central tendencies**: median, 242, 246; average, 247; mode, 247.
- Charters**, W. W., 15, 187, 221, 291.
- Childs**, H. G., 231, 239.
- Clark**, J. R., 227, 232, 275.
- Class instruction, waste of time in, 55; effects of, in arithmetic, 58; modified, 59.
- Cleveland Survey Arithmetic Tests**, 25; standards, 40-41; as instruments for diagnosis, 50 ff.
- Coefficient of correlation**, 251.
- Composition Scales**: Hillegas Scale, 194, 199, 200; Thorndike extension of Hillegas Scale, 195; Harvard-Newton Scale, 195, 200, 204; Breed and Frostic Scale, 195; Willing Scale, 196, 204, 206; Nassau County Supplement, 196.

- Composition, reliability of measurement in, 196 ff.
- Copying Test, 215 ff.; kinds of errors made, 217.
- Correlation, 250.
- Course of study, 289 ff.
- Courtis, S. A., 19, 37, 39, 46, 60, 73, 81, 120, 140.
- Courtis's English Tests, 73-74.
- Courtis Standard Research Tests, Series B, 23, 56; standards, 40; as instruments for diagnosis, 50.
- Courtis's Silent Reading Tests, 81-82, 89.
- Courtis's Standard Practice Tests in Arithmetic, 60; in Spelling, 140.
- Criteria for evaluating tests, 85-86.
- Culp, Vernon, 164.
- Cycle Principle, 275.
- Davidson, P. E., 111.
- Diagnosis:** in algebra, 232; in arithmetic, 32, 49 ff.; in handwriting, 158 ff.; in reading, 93 ff., 97; in spelling, 130-35.
- Efficiency, definition of, 264.
- Elliot, E. C., 6.
- Evaluation of exercises: per cent of pupils solving basis, 275; opinion of competent judges, 278; teacher-judgment, 279.
- Evaluation of reading tests, 85-90.
- Examinations:** sources of error in, 8 ff.; preparation of questions, 8 ff.; value of questions, 8 ff., 273 ff.; marking of papers, 5 ff., 280 ff.
- Example *vs.* problem, 22.
- Examples, types of, in arithmetic, 19.
- Fordyce, Charles, 120.
- Foreign languages:** Starch's tests, 234; Hanus Latin tests, 235; Hennmon Latin tests, 235.
- Fractions, tests in, 34; standards, 42.
- Freeman, F. N., 135, 136, 160, 161, 168, 169, 171, 172, 179, 296.
- Freeman's Handwriting Scale, 153-54; 160-62.
- Frostic, F. W., 194.
- Geometry**, test in, 233.
- Gilliland, A. R., 111.
- Grammar**, measurement of ability in, 212, 215.
- Grand Rapids, Michigan, scores for Cleveland Survey arithmetic tests, 41.
- Graphical representation, 253 ff.
- Graves, S. Monroe, 181.
- Gray, C. T., 154, 163, 167.
- Gray, W. S., 78, 83.
- Gray's Handwriting Score Card, 154-56, 158-59.
- Gray's Oral Reading Test, 83-85, 89.
- Gray's Silent Reading Tests, 78-79, 87, 89.
- Habit Formation:** laws of, applied to arithmetic, 53 ff.; to spelling, 135-36, 140; to handwriting, 183 ff.
- Haggerty, M. E., 69, 83, 288.
- Haggerty's Visual Vocabulary Tests, 83, 86.
- Handwriting:** problem of measurement, 145-46; securing specimens, 147-48; methods of using scales, 156-58; diagnosis, 158-62; using Freeman's scale, 161-62; reliability of scores, 162-64; training in using scales, 165-67; standards, 168-74; individual differences, 175 ff.; nature of ability, 175; remedial instruction, 180 ff.; systems of penmanship compared, 181; movement, 181; rhythm, 182, 186; speed, 182; laws of habit formation, 183 ff.; speed and quality, 183; devices of remedial instruction, 184 ff.; motivation of practice, 187.
- Handwriting Scales:** Thorndike, 148; Ayres, 152; Johnson and Stone, 152-53; Breed and Downs, 153; Freeman, 153-54; Gray, 154-56.
- Hanus, Paul, 235.
- Harvard-Newton Composition Scale, 195, 279; directions for using, 200; standards, 204.

- Harvey, Nathan A., 164.
 Hillegas, M. B., 194, 278.
Hillegas Composition Scale, 194,
 278; directions for using, 199;
 standards, 200.
 Houser, J. D., 143.
 Hudelson, Earl, 197.
 Hurt, A. O., 166.
- Individual differences**: arithmetic, 55 ff.; silent reading, 105; spelling, 116-17, 184-85; handwriting, 175 ff.
- Individual instruction**: arithmetic, 59 ff.; handwriting, 175; spelling, 130 ff.
- Inglis, Alexander, 8.
Intelligence Tests, 105.
- Johnson, F. W., 4, 222.
 Johnson, Harry, 164.
 Johnson, J. H., 167-69.
Johnson and Stone's Handwriting Scale, 152, 153.
 Jones, N. F., 114, 132-33.
 Jones, R. G., 82, 87.
Jones' Visual Vocabulary Tests, 82, 87, 89.
 Judd, C. H., 26, 171, 172.
- Kansas Silent Reading Tests, 79-81, 88-89, 91-92, 99.
 Kayfetz, Isidore, 194.
 Kelly, F. J., 4, 79, 163, 197, 280.
 King, Irving, 164.
 King, W. I., 246, 247.
- Lane, H. A., 298.
Language: problem of measurement, 192; measurement of ability by completion-tests, 210; value of tests, 218 ff.; analysis of ability, 220.
- Laws of habit formation**: applied to arithmetic, 53 ff.; applied to spelling, 135-36, 140; applied to handwriting, 183 ff.
- Lewis, E. E., 163, 174.
 Lull, H. G., 139.
- Manuel, H. T., 164, 191.
Median, 242, 246.
Minimum essentials, 15.
Minnesota Scale Beta, 73, 86-87.
Mode, 247.
 Monroe, Walter S., 37, 225, 228.
 Morrison, J. C., 301.
 Movement in handwriting, 181.
- Nassau County Supplement, 196.
 Nutt, H. W., 181.
- Oral Reading Tests**: Jones' visual vocabulary tests, 82; Haggerty's visual vocabulary tests, 83; Gray's Oral Reading Test, 83-85.
- Oral reading, overemphasis on, 99-100.
 Otis, A. S., 111, 117, 119.
Overlapping of grades, 56.
- Percentiles, 249.
Performance, *See Ability*.
Physics, test in, 237.
 Pinter, R., 111, 167.
 Practice tests in arithmetic, 60 ff.; in spelling, 140-41.
 Probable error, 249.
 Problem *vs.* example, 22.
- Problem of measurement**: arithmetic, 17 ff.; reading, 66; spelling, 112; handwriting, 145-46; language, 192; algebra, 224.
- Pupil, value of tests to, 95-96.
- Quartiles, 249.
- Rate of doing work neglected, 11.
- Reading**: problem of measurement, 66; diagnosis, 93-95, 97; remedial instruction, 96-107.
- Reading tests, *See Oral reading, or Silent reading*.
- Reading tests evaluated, 85, 90.
- Reasoning tests in arithmetic, 35.
- Reliability of measures**, 2; in arithmetic, 44 ff.; in reading, 87 ff.; in spelling, 116 ff.; in handwriting, 162 ff.; in composition, 196.
- Remedial instruction**, 287 ff.; in

- arithmetic, 53 ff.; in reading, 96 ff.; in spelling, 135 ff.; in handwriting, 180 ff.; in algebra, 231 ff.
- Rhythm, in handwriting, 182; development of, 186.
- Rice, J. M., 144.
- Richards, A. M., 111.
- Rugg, H. O., 224, 227, 232, 275.
- Sackett, L. F., 169.
- Salt Lake City, Utah, scores for Stone's Reasoning Test, 44.
- School marks**, inaccuracy of, 1 ff.; indefiniteness of, 258.
- School marks *vs.* scores, 259.
- Scientific management, principles of, 46 ff.; defined, 294 ff.
- Scores, accuracy of: arithmetic, 44 ff.; spelling, 116; handwriting, 162; composition, 196.
- Scores, translation of, 260 ff.
- Sears, J. B., 134.
- Silent Reading Tests**: Thorndike's Scale Alpha, 71-73; Minnesota Scale Beta, 73; Courtis's English Tests, 73-74; Brown's Silent Reading Test, 74-76; Starch's Silent Reading Tests, 76-78; Gray's Silent Reading Tests, 78-79; Kansas Silent Reading Tests, 79-81; Courtis's Silent Reading Tests, 81-82.
- Smith, James H., 53-56.
- Speed and quality, 183.
- Speed in handwriting, 182.
- Spelling demons, 132-33.
- Spelling**: definition of ability, 112-13, 125; problem of measurement, 112; making a spelling test, 114-24; individual differences, 116-17, 134-35; number of words to use, 118-19; a timed sentence test, 122-24; diagnosis, 130-35; types of errors, 133-35; remedial instruction, 135-42; causes of errors, 136-38; devices for teaching, 138-42; practice tests, 140-41.
- Standardized tests, use in supervision, 284 ff.; results of using, 298 ff.
- Standards**: Courtis Standard Re-
- search Tests, Series B, 40; Cleveland Survey Arithmetic Tests, 40-41; Woody's Arithmetic Scales, 42; Addition of Fractions, 42; Stone's Reasoning Test, 42 ff.; Thorndike's Visual Vocabulary Scale, A, 69; Thorndike's Scale, Alpha, 73; Brown's Silent Reading Test, 76; Starch's Silent Reading Tests, 78; Gray's Silent Reading Tests, 79; Kansas Silent Reading Tests, 81, 92, 99; Ayres's Spelling Scale, 128-30; Starch's Spelling Scale, 130; Handwriting Scales, 168-74; Hillegas's Composition Scale, 200; Harvard-Newton Composition Scale, 204; Trabue's Completion-Test Language Scales, 212; Starch's Grammatical Scales, 213; Copying Test, 217; Standard Research Tests in Algebra, 231.
- Standards**, basis of, 263 ff.; types of, 270; use of, 293 ff.
- Starch, Daniel, 6, 37, 70, 76, 125, 169, 171, 172, 212, 234, 237.
- Starch's Grammatical Scales, 212 ff.
- Starch's Silent Reading Tests, 76-78, 87, 88-89.
- Starch's Spelling Scales, 125-28, 130.
- Stockard, L. V., 233.
- Stone, C. R., 187.
- Stone, C. W., 18, 42.
- Stone's Reasoning Test, 35, 42.
- Studebaker, J. W., 62.
- Studebaker Economy Practice Exercises, 62.
- Superintendent, value of tests to, 90-93, 284 ff.
- Supervision, steps in, 285 ff.
- Taylor, F. W., 48.
- Teachers' Marks**, inaccuracy of, 2 ff.
- Teacher, value of tests to, 93-95.
- Teachers, specifications for, 288.
- Terman, L. M., 105.
- Tests, derivation of, 30, 34, 273 ff.
- Thorndike, E. L., 71, 118, 125, 148, 165, 194, 197, 246, 247, 249, 250, 251, 279.

- Thorndike's Extension of Hillegas Scale, 195.
Thorndike's Handwriting Scale, 148.
Thorndike's Reading Scale, Alpha, 71-73, 87.
Thorndike's Visual Vocabulary Scale, 67.
Tidyman, W. F., 125.
Trabue, M. R., 194, 197, 210, 275.
Trabue's Completion-Test Language Scale, 210 ff., 212.
Uhl, W. L., 95.
Variability, measures of: average deviation, 247; percentiles, quartiles, probable error, 249.
Vocabulary Scales; Thorndike's Scale A1, Scale A2, Scale B, 67-69; Haggerty's Scale, R2, 69; Starch, 70.
Wallin, J. E. W., 144.
Wassen, Alfred W., 90.
Willing, M. H., 194.
Willing's Composition Scale, 196; directions for using, 204; standards, 206; scale reproduced, 206 ff.
Wilson, G. M., 172, 187.
Wilson, H. B., 187.
Witham, E. C., 172.
Woody, Clifford, 29, 275.
Woody's Arithmetic Scales, 29, 42; as an instrument for diagnosis, 33, 53.
Ziedler, Richard, 111.

RIVERSIDE TEXTBOOKS IN EDUCATION

HEALTHFUL SCHOOLS: HOW TO BUILD, EQUIP, AND MAINTAIN THEM.

By MAY AYRES, JESSE F. WILLIAMS, M. D., Professor of Hygiene and Physical Education, University of Cincinnati, and THOMAS D. WOOD, A. M., M. D., Professor of Physical Education, Teachers College, Columbia University. \$1.50 net. Postpaid.

TEACHING LITERATURE IN THE GRAMMAR GRADES AND HIGH SCHOOL.

By EMMA M. BOLENIUS, formerly Instructor in English, Central Commercial and Manual Training High School, Newark, N. J. \$1.50 net. Postpaid.

PUBLIC EDUCATION IN THE UNITED STATES.

By ELLWOOD P. CUBBERLEY. *In preparation.*

PUBLIC SCHOOL ADMINISTRATION.

By E. P. CUBBERLEY. \$2.00 net. Postpaid.

RURAL LIFE AND EDUCATION.

By ELLWOOD P. CUBBERLEY, Dean of the School of Education, Leland Stanford Junior University. \$1.60 net. Postpaid.

EVOLUTION OF THE EDUCATIONAL IDEAL.

By MABEL I. EMERSON, First Assistant in Charge of the George Bancroft School, Boston. \$1.20 net. Postpaid.

EXPERIMENTAL EDUCATION.

By F. N. FREEMAN. \$1.50 net. Postpaid.

HOW CHILDREN LEARN.

By F. N. FREEMAN. \$1.60 net. Postpaid.

THE PSYCHOLOGY OF THE COMMON BRANCHES.

By F. N. FREEMAN, Assistant Professor of Educational Psychology, University of Chicago. \$1.50 net. Postpaid.

HEALTH WORK IN THE SCHOOLS.

By E. B. HOAG, M. D., Medical Director, Long Beach City Schools, California, and L. M. TERMAN. \$1.75 net. Postpaid.

HOW TO TEACH THE FUNDAMENTAL SUBJECTS.

By C. N. KENDALL, Commissioner of Education for New Jersey, and G. A. MIRICK, formerly Deputy Commissioner of Education for New Jersey. \$1.50 net. Postpaid.

HOW TO TEACH THE SPECIAL SUBJECTS.

By C. N. KENDALL and G. A. MIRICK. \$1.60 net. Postpaid.

EDUCATIONAL TESTS AND MEASUREMENTS.

By W. S. MONROE, Director of the Bureau of Coöperative Research, University of Indiana; J. C. DeVoss, Associate Professor of Psychology and Philosophy, Kansas State Normal School; and F. J. KELLY, Dean of the School of Education, University of Kansas. \$1.60 net. Postpaid.

MEASURING THE RESULTS OF TEACHING.

By W. S. MONROE. \$1.60 net. Postpaid.

DISCIPLINE AS A SCHOOL PROBLEM.

By A. C. PERRY, Jr., District Superintendent of Schools, New York City. \$1.50 net. Postpaid.

STATISTICAL METHODS APPLIED TO EDUCATION.

By H. O. RUGG, Assistant Professor of Education, University of Chicago. \$2.00 net. Postpaid.

CLASSROOM ORGANIZATION AND CONTROL.

By J. B. SEARS, Associate Professor of Education, Leland Stanford Junior University. \$1.75 net. Postpaid.

AN INTRODUCTION TO EDUCATIONAL SOCIOLOGY.

By WALTER R. SMITH, Professor of Sociology and Economics, Kansas State Normal School, Emporia, Kansas. \$1.75 net. Postpaid.

THE HYGIENE OF THE SCHOOL CHILD.

By L. M. TERMAN, Professor of Education, Leland Stanford Junior University. \$1.75 net. Postpaid.

THE INTELLIGENCE OF SCHOOL CHILDREN.

By L. M. TERMAN. *In preparation.*

THE MEASUREMENT OF INTELLIGENCE.

By L. M. TERMAN. \$1.75 net. Postpaid.

Test Material for the Measurement of Intelligence. 60 cents net. Postpaid. Record Booklets (in packages of twenty-five). \$2.00 net, a package. Postpaid.

THE TEACHING OF SCIENCE IN THE ELEMENTARY SCHOOL.

By GILBERT H. TRAFTON, Instructor in Science, State Normal School, Mankato, Minnesota. \$1.30 net. Postpaid.

AN INTRODUCTION TO CHILD PSYCHOLOGY.

By CHARLES W. WADDLE, Ph. D., Supervisor of Practice Teaching, Los Angeles State Normal School. \$1.50 net. Postpaid.

TEACHING IN RURAL SCHOOLS.

By T. J. WOOFTER, Dean of the School of Education, University of Georgia. \$1.40 net. Postpaid.

Secondary Education Division

PRINCIPLES OF SECONDARY EDUCATION.

By ALEXANDER INGLIS, Assistant Professor of Education, Harvard University. \$2.75.

PROBLEMS OF SECONDARY EDUCATION.

By DAVID SNEDDEN, Professor of Education, Teachers College, Columbia University. \$1.50 net. Postpaid.

THE TEACHING OF ENGLISH IN THE SECONDARY SCHOOLS.

By CHARLES SWAIN THOMAS, Head of the English Department, Newton High School, Newton, Mass. \$1.60 net. Postpaid.

HOUGHTON MIFFLIN COMPANY
BOSTON NEW YORK CHICAGO

June 24 Corn Shpt.

